

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
23 April 2009 (23.04.2009)

PCT

(10) International Publication Number
WO 2009/052277 A1

- (51) International Patent Classification:
G06F 17/20 (2006.01)
- (21) International Application Number:
PCT/US2008/080149
- (22) International Filing Date: 16 October 2008 (16.10.2008)
- (25) Filing Language: English
- (26) Publication Language: English
- (30) Priority Data:
60/980,747 17 October 2007 (17.10.2007) US
- (71) Applicant (for all designated States except US): EVRI, INC. [US/US]; 206 - 1st Avenue South, Suite 310, Seattle, Washington 98104 (US).
- (72) Inventors; and
- (75) Inventors/Applicants (for US only): LIANG, Jisheng [CN/US]; 6343 - 114th Avenue Southeast, Bellevue, Washington 98006 (US). KOPERSKI, Krzysztof [CA/US]; 2311 Yale Avenue East, Apt. D, Seattle, Washington 98102 (US). DHILLON, Navdeep, S. [US/US]; 8011 - 29th Avenue Northwest, Seattle, Washington 98117 (US). TUSK, Carston [DE/US]; 20912 - 4th Avenue South, Seattle, Washington 98198 (US). BHATTI, Satish [US/US]; 523 North 101 Street, Seattle, Washington 98133 (US).
- (74) Agents: BIERMAN, Ellen, M. et al.; Seed Intellectual Property Law Group PLLC, Suite 5400, 701 Fifth Avenue, Seattle, Washington 98104-7064 (US).
- (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, BR, BW, BY, BZ, CA, CH, 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, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, 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, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, NO, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

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

[Continued on next page]

(54) Title: NLP-BASED ENTITY RECOGNITION AND DISAMBIGUATION

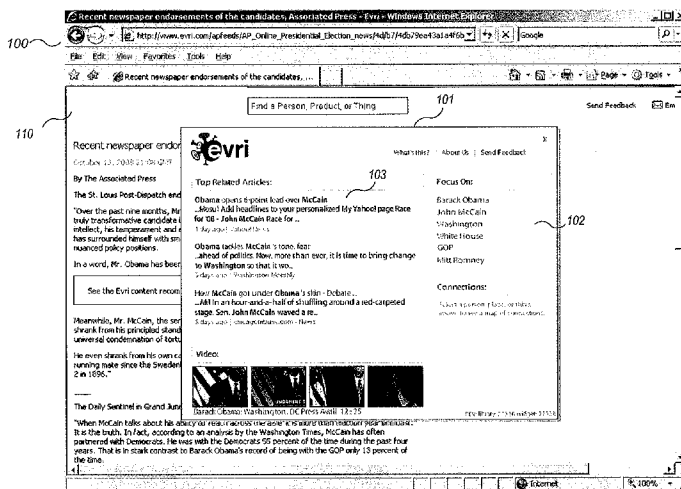


Fig. 1

(57) Abstract: Methods and systems for entity recognition and disambiguation using natural language processing techniques are provided. Example embodiments provide an entity recognition and disambiguation system (ERDS) and process that, based upon input of a text segment, automatically determines which entities are being referred to by the text using both natural language processing techniques and analysis of information gleaned from contextual data in the surrounding text. In at least some embodiments, supplemental or related information that can be used to assist in the recognition and/or disambiguation process can be retrieved from knowledge repositories such as an ontology knowledge base. In one embodiment, the ERDS comprises a linguistic analysis engine, a knowledge analysis engine, and a disambiguation engine that cooperate to identify candidate entities from a knowledge repository and determine which of the candidates best matches the one or more detected entities in a text segment using context information.

WO 2009/052277 A1



— *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))*

Published:

— *with international search report*

NLP-BASED ENTITY RECOGNITION AND DISAMBIGUATION

TECHNICAL FIELD

The present disclosure relates to methods, techniques, and systems for entity identification using natural language processing and, in particular, to methods and systems for recognizing and disambiguating named entities using natural language processing, knowledge repositories, and/or other contextual information.

BRIEF SUMMARY

Embodiments described herein provide enhanced computer- and network-assisted methods, techniques, and systems for recognizing and disambiguating entities in arbitrary text using natural language processing. Example embodiments provide an entity recognition and disambiguation system (an "ERDS") and process that, based upon input of a text segment, automatically determines which entities are being referred to by the text using both natural language processing techniques and analysis of information gleaned from contextual data in the surrounding text. In at least some embodiments, supplemental or related information that can be used to assist in the recognition and/or disambiguation process can be retrieved from knowledge repositories such as an ontology knowledge base.

According to one approach, a method, is provided that identifies one or more entities in an indicated text segment by processing the indicated text segment; recognizes one or more candidate named entities which are referred to by a detected entity in a received text segment based, at least in part, upon a natural language analysis of the text segment; disambiguates the candidate named entities to determine a single named entity to which the detected entity in the received text segment is deemed to refer based upon a combination of one or more of linguistic analysis, contextual information gleaned from surrounding text, or information retrieved from one or more knowledge repositories.

Computing systems and computer-readable media containing content that performs a portion or all of the methods described above are also provided.

BACKGROUND

With the proliferation of information generated daily and accessible to users over the Web, the need for intelligent electronic assistants to aid in locating and/or discovering useful or desired information amongst the morass of data is paramount. The use of natural language processing to search text to correctly recognize people, places, or things is fraught with difficulties. First, natural language is ambiguous. Almost every English word or phrase can be a place name somewhere in the world or a name of a person (*i.e.*, a "person name"). Furthermore, many entities share the same name. For example, there are more than 20 cities named "Paris" in the United States. A person named "Will Smith," could refer to the Hollywood movie actor and musician, the professional football player in the NFL, or many other people. Recognizing non-celebrity names has become more important with the exponential growth of the Web content, especially user created content such as blogs, Wikipedia, and profiles on social network sites like MySpace and FaceBook. Second, an entity could be mentioned or referred to in many different ways (*e.g.*, pronouns, synonyms, aliases, acronyms, spelling variations, nicknames, etc.) in a document. Third, various knowledge sources about entities (*e.g.* dictionary, encyclopedia, Wikipedia, gazetteer, etc.) exist, and the size of these knowledge bases are extensive (*e.g.* millions of person names and place names). The sheer quantity of data is prohibitive for many natural language processing techniques.

BRIEF DESCRIPTION OF THE DRAWINGS

The patent or application file contains at least one drawing executed in color. Copies of this patent or patent application publication with color drawings will be provided by the Office upon request and payment of the necessary fee.

Figure 1 is an example screen display of an example pop-up window that incorporates ERDS functionality to identify entities in an underlying news article.

Figure 2 is an example screen display of further use of an identified entity to obtain additional information.

Figure 3 is an example block diagram of components of an example Entity Recognition and Disambiguation System.

Figure 4 is an example block diagram of an overview of an example entity recognition and disambiguation process used by an example embodiment of an Entity Recognition and Disambiguation System.

Figure 5 is an example flow diagram of the sub-steps of the entity recognition process based upon linguistic analysis of the input text.

Figure 6 is an example illustration of SAO triplets along with descriptive modifiers.

5 Figure 7 is a schematic illustration of long-distance dependency recognition.

Figure 8 is an example flow diagram of an example disambiguation process of an example embodiment.

10 Figure 9 is an example block diagram of an example computing system that may be used to practice embodiments of a recognition and disambiguation system such as that described herein.

DETAILED DESCRIPTION

Embodiments described herein provide enhanced computer- and network-assisted methods, techniques, and systems for recognizing and disambiguating entities in arbitrary text using natural language processing.

5 Example embodiments provide an entity recognition and disambiguation system (an "ERDS") and process that, based upon input of a text segment, automatically determines (*e.g.*, makes a best analysis of identifying) which entities are being referred to by the text using both natural language processing techniques and analysis of information gleaned from contextual data in the surrounding text. The
10 text segment may comprise a portion or all of a document, and may be part of a larger composite collection. In at least some embodiments, supplemental or related information that can be used to assist in the recognition and/or disambiguation process can be retrieved from knowledge repositories such as an ontology knowledge base.

15 Entities refer generally herein to people, places, or things. Entities are determined based upon recognizing mentions of entities and disambiguating those that potentially correspond to more than one entity. Entity disambiguation refers to the process for differentiating which entity is being referred to in a given piece of text when a recognized entity may correspond to more than one entity.
20 The recognition and disambiguation process may further set up a correspondence between the determined entities and their corresponding entries stored in a knowledge repository and return a list of identified entities. In addition, the identified entities may be ranked and/or grouped based on their relevance/importance to the input text.

25 Once the entities in the input text are identified, the results of the recognition and disambiguation process can be used to present useful information such as to summarize information in an underlying Web page, to present supplemental related data, etc. For example, a listed of ordered or ranked entities can be provided along side information presented for other purposes, such as in
30 conjunction with other information presented to a user while perusing web pages to research or otherwise discover useful information. In addition, links to other related information can be provided based upon the knowledge acquired by the ERDS while performing entity identification. For example, the coordinates of place names could be retrieved, and, as a result the recognized place names could be
35 displayed on a map. Also, the identified entities in a document (*e.g.*, their links) could be automatically highlighted so that further information on the entities could

be easily accessed (e.g., the links could automatically point to knowledge source entries). Further, better summarization of document content could be provided, for example, by listing the “top” ranked entities in juxtaposition to the document, targeting advertising toward the top ranked entities, etc. A variety of other uses
5 could be similarly employed to integrate the results of entity recognition and disambiguation techniques.

The ERDS (or entity recognition and disambiguation process) may be made available as a standalone module such as a widget, pop-up window, etc., embedded in other programs or modules, provided as a service, accessed via an
10 application programming interface (“API”), etc. In other contexts, the recognition and disambiguation process used in an ERDS may be incorporated in powerful indexing and search tools, such as those used to provide relationship searches that find information based upon natural language analysis rather than merely keyword searching.

15 Figure 1 is an example screen display of an example pop-up window that incorporates ERDS functionality to identify entities in an underlying news article. In Figure 1, using a Web browser 100, a user has navigated to a page 110 that displays a recent newspaper article from Associated Press, entitled “Recent newspaper endorsements of the candidates.” The user has selected (for example,
20 by means of an input device such as a mouse) an icon (covered up) that provides a link to a widget, here a pop-up window 101, that incorporates ERDS functionality to recognize and disambiguate where needed the topmost entities described in the article. In pop-up window 101, the list of recognized entities 102 is provided as possible further focuses for the user. The list 102 is shown ranked in an order. In
25 some embodiments, the ranked order may be based upon a frequency that is not just based on the article, but possibly based upon other references to the entities (e.g., in other news articles, frequency of user request, etc.) elsewhere in the universe being examined. The Top Related Articles section 103 provides links to other articles relating to the recognized entities shown in the list 102. These other
30 articles may be ordered based upon factors such as popularity and what is most current.

Figure 2 is an example screen display of further use of an identified entity to obtain additional information. In this case, the user has selected one of the recognized entities, the John McCain entity 203. In response, the widget 201
35 that incorporates the ERDS techniques presents a connections diagram 204 of other entities that the selected entity 203 is connected to. In addition, the Top Related Articles section 205 is modified to reflect the changed entity selection.

Further selection of one of the connected entities in connections diagram 204, such as connected entity 206, can further cause the articles shown in section 205 to change. Accordingly, widgets 101 and 201 may be used to perform research, discover further useful information, etc, without the user explicitly needing to enter
5 search commands or keywords in a search language.

Additional example embodiments that include different user interfaces for presenting and manipulating determined entities and/or the information connected to such entities are described in co-pending U.S. Patent Application (Attorney Docket No. 470064.405), entitled NLP-BASED CONTENT
10 RECOMMENDER, filed concurrently herewith, and herein incorporated by reference in its entirety.

Figure 3 is an example block diagram of components of an example Entity Recognition and Disambiguation System. In one embodiment, the Entity Recognition and Disambiguation System comprises one or more functional
15 components/modules that work together to automatically determine which entities are referred to in specified text (e.g., from input documents). These components may be implemented in software or hardware or a combination of both.

In Figure 3, the Entity Recognition and Disambiguation System 301 comprises a linguistic analysis engine 311, a knowledge analysis engine 312, a disambiguation engine 313, a document data repository 316, and a knowledge
20 data repository 315. The ERDS 301 may optionally comprise an NLP parsing/preprocessor, potentially as a separate component, especially in instances where a third party or off-the-shelf processor is integrated into the system. The ERDS analyzes text segments from text documents 310 to automatically
25 determine the various entities contained in them, and potentially outputs entity identifiers 320.

In some embodiments, the entity identifiers 320 are stored as links in the text documents 310 and setup correspondences with entity information stored in one or more knowledge repositories 315, e.g., ontology knowledge bases. In
30 addition, in some embodiments, the entity information is stored and indexed in the document data repository 316, for example to assist in performance of relationship searching, along with other data indexed from the documents 310. This is particularly useful in embodiments that embed the ERDS as part of a larger document storage and indexing system for searching. The linguistics analysis
35 engine 311 receives text from the text documents 310 which has been potentially preprocessed by preprocessor 314 and determines characteristics about the entities mentioned (but not yet resolved) in the documents 310. The knowledge

analysis engine 312 matches these characters to find possible candidate entities in a vast array of knowledge information, represented in part by the knowledge data 315. The disambiguation engine 315 resolves which possible candidate entity is the most likely entity being referred to in the text, and returns indications of the
5 identified (resolved) entities.

One or more embodiments of an ERDS may incorporate an existing natural language processing technology and relationship search technology, called InFact®. The InFact® NLP-based relationship search technology, now referred to as the Evri relationship search technology, is described in detail in U.S. Patent
10 Application No. 11/012,089, filed December 13, 2004, which published on December 1, 2005 as U.S. Patent Publication No. 2005/0267871A1, and which is hereby incorporated by reference in its entirety. The InFact®/Evri relationship search and indexing technology defines data structures, techniques, and systems for natural language processing that preprocess and store document information in
15 one or more annotated phrase, clause, or sentence-based indexes. The annotations store tag-based information, along with the terms of a clause or sentence, so that existing keyword-based search engines can be leveraged to perform natural language based processing instead of simple keyword searching. The tag-based information may include one or more of parts-of-speech tags,
20 grammatical role tags, ontological information, entity tags, and other types of information. The InFact®/Evri relationship search technology also defines methods, systems, and techniques for performing relationship queries, as opposed to traditional keyword-based searching. These relationship queries leverage NLP to allow users to discover information regarding how a particular entity relates or is
25 otherwise associated with other entities. Accordingly, the recognition and disambiguation process used in an ERDS also may be incorporated into an InFact®/Evri text analysis and indexing system, and can be invoked during the indexing of documents to identify and store entity related information.

Figure 4 is an example block diagram of an overview of an example
30 entity recognition and disambiguation process used by an example embodiment of an entity recognition and disambiguation system. The ERDS techniques addresses some of the challenges posed by recognizing entities, for example within the voluminous amounts of information promulgated by the Web, by disambiguation based upon InFact®/Evri NLP techniques, normalization of
35 variations of languages, especially the resolution of coreferences of same entities, and development of an ontology representation and repository that provides fast and scalable access to knowledge sources about entities.

In block 401, the process performs natural language parsing and linguistic analysis (preprocessing) of each sentence in the input document to turn the text into structured information. The term "document" as used herein refers to any portion or all of one or more text segments.

5 In block 402, it detects/recognizes potential entity names by combining evidence from multiple levels, based on a more sophisticated syntactic and semantic analysis as opposed to simple keyword-based matching. As the result, a profile is built for each entity mentioned in the text that characterizes the entity's properties based upon the document content.

10 Then, in block 403, the process performs an ontology lookup on each potential entity name using an ontology lookup service. The ontology lookup service provides the capability, given a name and its attributes, to retrieve all matching ontology entries. To perform this processing, for each detected entity, a search is performed against the ontology repository, which is likely to return
15 multiple matches (or candidate entries in the ontology repository that potentially match the name/entity profile). The result of this processing is one or more candidate entities.

In block 404, the process then performs an entity disambiguation process for the candidate entries. The entity disambiguation step determines the
20 correct entity that the text is likely referring to, from among the multiple candidate entries in the ontology (or determines that none of the candidate entries is correct). This disambiguation process is performed by examining the candidate's attributes specified in the ontology and comparing them to contextual information extracted from the text (*i.e.*, the profiles derived from the text). What results is one or more
25 identified entities.

Then, in block 405, the identified (disambiguated) entities are ranked based on certain criteria (*e.g.*, frequency of occurrence in the document) or grouped based on their common characteristics (*e.g.*, places that belong to same region or people with the same profession). The entity recognition and
30 disambiguation process for the received input is then complete.

In addition, in some embodiments that integrate these techniques with InFact®/Evri relationship search and indexing technology, in block 406, the matching ontology entries (to the identified entities) are stored, as annotations of phrases within the sentence-level indices, to support queries against them. (They
35 may also be stored at other level indices in some other embodiments.) The InFact®/Evri search and indexing technology supports queries of the annotated entities and their relationships. To do so, the indexing and storage mechanism

stores each sentence of a document as a series of clauses, annotated by one or more tags, wherein the tags may include entity tag information, grammatical role information, syntactic (parts-of-speech) roles, etc. The structure and use of these inverted indexes is described further in U.S. Patent Publication No. 5 2005/0267871A1.

As illustrated in Figure 4, the entity recognition and disambiguation process generates several intermediate results in the process of identifying entities. The following example is illustrative:

Example text:

10 Carolina quarterback David Carr was sacked by Saints defensive end Will Smith in the first quarter. Smith leads the Saints in sacks. He was a Pro-Bowl starter last year.

15 The following Table 1 shows the result of processing after the linguistic analysis:

Subject : Modifier	Action : Modifier	Object: Modifier
Will Smith [Person name] : defensive end, Saints	sack : in first quarter	David Carr [Person name]: Carolina, quarterback
Will Smith [Person name]	lead : in sack	Saints [Organization]
Will Smith [Person name]	be	starter: Pro-Bowl, last year

Table 1- Linguistic Analysis

20 In Table 1, note that Will Smith is tagged as a person name. The three coreferences (“Will Smith”, “He”, and “Smith”) are resolved and linked together so that they identify the same entity.

Next (after block 402), an entity profile is constructed for Will Smith. The entity profile describes the characteristics of the entity as specified in the text. Table 2 shows the constructed entity profile.

Actions	Modifiers	Related Entities
sack, lead	defensive end, starter	Saints, Carolina, David Carr, Pro-Bowl

Table 2 – Entity Profile

This entity profile is then looked up in the ontology data repositories. Using the name “Will Smith” as query, the process finds three matches (candidate entities) from the ontology repository (after block 403) as shown below in Table 3. Each entry in the ontology specifies a number of attributes (*i.e.*, type, role, related entities). Other attributes may be available.

Name	Type/Role	Related Entities
Will Smith	Person / Movie Actor	Independence Day, Men in Black, Ali,
Will Smith	Person / Football Player	New Orleans Saints, NFL, Ohio State University
Will Smith	Person / Cricket player	England, Nottinghamshire

Table 3 – Candidate Entities

Finally, the candidate entities are disambiguated. The disambiguation process (block 404) matches Will Smith's profile collected from the text against the three ontology matching entries, and determines that Will Smith the football player is the correct entity the text refers to. Multiple algorithms are available to make this determination. A few of them are described further below. The following identified entity shown in Table 4 results.

Will Smith	Person / Football Player	New Orleans Saints, NFL, Ohio State University
------------	--------------------------	--

Table 4 – Disambiguation Result (Identified Entity)

Example embodiments described herein provide applications, tools, data structures and other support to implement an entity recognition and disambiguation system and/or process to be used for assisting in the locating and

understanding of relevant information. Other embodiments of the described techniques may be used for other purposes.

In the following description, numerous specific details of one example embodiment are set forth, such as data formats and code sequences, etc., in order to provide a thorough understanding of the described techniques. The described approaches and algorithms may be used alone or in combination to implement an NLP-based recognition and disambiguation system. In some embodiments both natural language processing and knowledge based approaches are used. In other embodiments only some of the approaches or capabilities are incorporated. The embodiments described also can be practiced without some of the specific details described herein, or with other specific details, such as changes with respect to the ordering of the code flow, different code flows, etc. Thus, the scope of the techniques and/or functions described are not limited by the particular order, selection, or decomposition of steps described with reference to any particular routine.

Also, although certain terms are used primarily herein, other terms could be used interchangeably to yield equivalent embodiments and examples. For example, it is well-known that equivalent terms in the natural language processing field and in other similar fields could be substituted for such terms as "document," "text," etc. In addition, terms may have alternate spellings which may or may not be explicitly mentioned, and all such variations of terms are intended to be included.

As described in Figures 3 and 4, a primary function of an ERDS is to recognize and disambiguate entities present in specified text. This function may be provided, for example, as part of an indexing and storage process for performing relationship searches, such as available through the InFact®/Evri NLP relationship search technology. Once the entities are "resolved" (recognized and disambiguated), appropriate indications of the entities (e.g., entity tags) may be stored in the indexes that represent the clause/sentence/document structures, so that matches against those entities may be found efficiently.

Additional detailed information on the InFact®/Evri indexing and storage techniques and use of them for searching structured and unstructured data is available in several patents, publications and patent applications, including U.S. Patent No. 7,398,201; U.S. Patent Publication No. 2005/0267871A1; U.S. Patent Publication No. 2007/0156669A1; and U.S. Patent Application No. 12/049,184, filed March 14, 2008, which are incorporated herein by reference in their entireties.

The following sections describe each aspect of the entity recognition and disambiguation process and ERDS component interactions in more detail.

Pre-Processing

As mentioned in block 401 of Figure 4, the input text segment (document, etc.) is pre-processed, for example, using the NLP parsing/preprocessor 314 of Figure 3. The preprocessor takes input text and produces one or more data structures that contain the structure of regions (e.g., paragraphs), sentences, as well as associated meta-tags (e.g., publisher, URL, topic categories such as Politics, Entertainment, etc.)

10 Entity Recognition based upon Linguistic Analysis

Once pre-processed, the resulting structures are sent to the linguistic analysis engine, for example engine 311 of Figure 3, to recognize all of the entities in the inputted text (see, e.g., block 402 of Figure 4). Figure 5 is an example flow diagram of the steps of the entity recognition process based upon linguistic analysis of the input text. The entity recognition process detects entities mentioned in the input text, and links together multiple occurrences of same entity. The process consists of three sub-steps. In step 501, syntactic analysis of sentences (e.g., producing a parse tree of each sentence) is performed. In step 502, the process detects entity occurrences in the syntactic structure resulting from step 501 (the parse tree). In step 503, the process resolves coreferences of the same entity, for example determines which pronouns refer to which entities. Each of these steps is described in more detail next.

Syntactic Analysis

Given a sentence (or clause), a syntactic parser (e.g., the InFact®/Evri dependency parser) is used to produce a parse tree. A parse tree represents the syntax of a sentence as a tree structure and typically captures the following information:

- Part-of-speech tags (e.g., noun, adjective, preposition, verb, etc.)
- Grammatical roles (e.g., subject, object, verb, etc.)
- Multi-word Phrase structures, such as noun phrases

When used with a dependency parser (such as the InFact®/Evri dependency parser), the produced parse tree is a dependency tree, which is a type of parse tree representation. It represents the whole parse tree as a list of head-modifier relations (e.g., verb-subject relations, adjective-noun modifier

relations, etc.). As a result, each sentence can be decomposed into one or more of the subject-action-object triples (SAO triplets), along with their descriptive modifiers. Figure 6 is an example illustration of SAO triplets along with descriptive modifiers. SAO triplet 600 contains a subject 601, an action 602, and an object
5 603, along with their respective modifiers 604, 605, and 606.

The descriptive modifiers 604 and 606 include immediate modifiers of a noun, including appositive, nominative, genitive, prepositional, adjective modifiers, as well as normalized long distance dependencies, e.g., predicate nominative (the is-a relation). The subject-action-object event triplets indicates
10 entities' roles in an event (as source or target) and their interactions with other entities.

One of the benefits of using the InFact®/Evri dependency tree parsing is that it explicitly handles long-distance dependency in a normalized way. Figure 7 is a schematic illustration of long-distance dependency recognition. In the
15 examples shown in Figure 7, in clause 701 "president" is identified as direct modifier of head noun "Washington," even though there are many words between the pair of words. Similarly, in clause 702 "city" is a direct modifier of the head noun "Alexandria." These kind of long-distance descriptive modifier relations are difficult to capture by methods based on a finite-state model, such as regular
20 expression, n-gram language model, or Hidden Markov Model, which represent sentence structure as a sequence of words. Accordingly, use of the InFact®/Evri dependency tree parsing improves accuracy of entity recognition.

Entity Detection of Proper Nouns

Using the phrases and syntactic structures (the SAO triplets)
25 produced by the previous stage (step 501 of Figure 5), the entity recognition process then performs detection of proper nouns by examining evidence from multiple levels, some examples of which are described here. Proper nouns (also called proper names) are nouns representing unique entities (such as "London" or "John"), as distinguished from common nouns which describe a class of entities
30 (such as city or person). This entity detection/recognition step detects the proper nouns and assigns them into a number of general categories (e.g., people, place, organization, date, number, money amount, etc.). Other categories could be similarly supported.

More specifically, this step of entity recognition first rules out words
35 or phrases with part-of-speech tags other than noun, such as verb, preposition, and adjective. Many words have multiple senses (e.g., a word could be used as a

verb or noun). The recognition process resolves such ambiguity through part-of-speech tagging during the sentence parsing process.

Proper names are usually capitalized in English. They also have certain internal structures (e.g., a person name may have the structure of given name followed by family name). The entity detection step (step 502) applies lexical analysis to the words and phrases (from the produced dependency tree) that have not been ruled out to detect proper names from phrases. The analysis is based on a combination of dictionary entries and a set of regular expression patterns. Examples of such patterns that are recognized include:

- 10 • A set of common given names and family names can be collected in the dictionary. Then, if a recognized phrase (e.g., "George Vancouver") matches a pattern such as <given name> <capitalized word>, the entity recognition process can tag the phrase as a person name.
- 15 • A set of geographic keywords indicating place names (e.g., lake, mountain, island) can be collected in the dictionary. Then, the entity recognition process can recognize patterns such as <capitalized word> <geographic keyword> to identify place names (e.g., "Vancouver Island").
- 20 • Similarly, organization names can be identified by organization-related keywords appearing in the names (e.g., recognition of the keyword "University" in the phrase "University of Washington").

Furthermore, the entity recognition process can recognize certain types of proper names based on their modifiers, by detecting occurrence of the above mentioned keyword indicators. For example, title words (e.g., senator, congressman, president) present in modifiers usually indicate the head noun is a person name.

In contrast, the examples below illustrate that the word "Sofia" can be tagged as a city based on its modifiers, instead of as a person name:

- Appositive modifier: "Sofia, capital city of Bulgaria"
- Predicate modifier: "Sofia is the capital of Bulgaria"
- Nominal modifier: "Bulgarian capital Sofia"

Names appearing together through conjunctions (e.g., "and," "or") or as a list, more than likely belong to the same entity type. For example, if an unknown name appears together with a list of person names, the entity recognition

process labels the unknown name as a person name. Similarly, given a list of names “Moses Lake, George, Wenatchee,” the entity recognition process determines that “George” is likely a place name rather than a person name based on the recognition that its conjunctions are place names.

5 Coreference Resolution

An entity is often referred to in many different ways and at different locations in a segment of text. The entity recognition process applies a document-level coreference resolution procedure (step 503) to link together multiple occurrences (mentions) that designate the same entity.

10 The types of coreferences handled may include one or more of:

- Names and aliases:

In a given document, an unambiguous name is often mentioned in the beginning, then some form of acronyms or aliases are used in the rest of the document. These aliases are often localized, and may not be available from a domain lexicon or have ambiguous meanings. For example, “Alaska” refers to
15 “Alaska Airlines;” “Portland” refers to “Portland, Maine;” and “UT” refers to “University of Texas” instead of the state of “Utah.”

- Definite noun phrase anaphora (e.g., “Washington” is referred later as “the state”); and

20 • Pronoun anaphora (it, they, which, where, etc.).

For each potential anaphora (e.g., each pronoun), the entity recognition process determines its antecedent (the entity the anaphora refers to) using the following steps:

1. Determine the proper context for this kind of anaphora. For
25 example, the context for a pronoun is usually no more than two sentences.

2. Inspect the context surrounding the anaphora (e.g., pronoun) for candidate antecedents (nouns and the entities to which they refer) that satisfy a set of consistency restrictions. For example, if the current pronoun is “she” then only the Person entities with gender “female” or “unknown” will be considered as
30 candidate entities. Each antecedent noun at this point has a corresponding entity associated with it from the other analysis sub-steps.

3. Assign scores to each antecedent based on a set of weighted preferences (e.g., distance between the antecedent and the anaphora in the parse tree).

4. Choose as the candidate entity for the pronoun, the entity that corresponds to the antecedent noun with the highest score.

After performing syntactic analysis, entity detection of proper nouns , and resolving coreferences, the entity recognition process has determined to which
 5 named entities, the text probably refers. These are referred to as recognized entities or recognized named entities. It will be understood that variations of these steps could also be supported, such as different rules for determining which antecedent is likely the best candidate entity.

Building Entity Profiles

10 Again referring to Figure 4, once one or more named entities have been recognized in the text, corresponding entity profiles are built (as part of block 402) by processing each named entity in relation to the document to determine characteristics for each named entity. It is from these characteristics and other information that the entity recognition and disambiguation process (hence the
 15 ERDS) can decipher which precise entity is being referred to. In one example embodiment, for each entity, the ERDS builds an entity profile of the named entity by collecting the following information by examining each mention (direct, e.g., by name, or indirect, e.g., through a pronoun) of the entity in the document:

- 20 • list of descriptive modifiers of the entity (e.g. "Will Smith, the football player)
- list of actions where the entity is a subject (e.g. "Will Smith starred in the movie)
- list of actions where the entity is a object (e.g. "arrive at JFK")
- 25 • list of other entities that interact with the given entity (e.g. "Will Smith plays for the New Orleans Saints")
- collection of general categories assigned to the entity (i.e. Person, Location, Organization), for example, by the InFact®/Evri dependency parser.

30 In other embodiments less or more information may be collected for each mention of an entity.

Once the context information is collected, the ERDS further assigns more specific roles (e.g., politician, musical artist, movie actor, tennis player, etc.) to each entity profile. There are several ways to gain additional information that

assists with assigning roles. For example, title words such as “mayor”, “senator”, “presidential candidate” appearing as descriptive modifiers likely indicate that the entity has been used in a “politician” role. Roles can also be indicated by the actions that entities are involved in. For example, in the clause “JFK was assassinated,” JFK is a person name, while in the clause “arrive at JFK,” JFK is more likely a place name.

In addition, the collected information for an entity can be compared to known roles in order to match the entity to its most likely role (*i.e.*, category). Table 5 below illustrates two clusters of people based on their action patterns collected from a set of news articles.

person name with similar actions	common action profiles
steve young, joe montana, jim everett, jim mcmahon, steve bono, troy aikman, john elway, phil simms, mike pawlawski, mark vlastic	throw, play, complete, get, start, lob, finish, share, run, sustain, wear, seem, come in, take, recover, rally, scramble, spot, find, make, convert, lead, game, replace, pick up
andre agassi, steffi graf, john mcenroe, monica seles, martina navratilova, boris becker, ivan lendl, michael chang, stefan edberg, pete sampras, gabriela sabatini, jimmy connors, chris evert, jennifer capriati, becker, jim courier, andres gomez	play, beat, win, defeat, lose, overcome, reach, wear, route, take, rally, withdraw, skip, serve, get, come, make, hit, miss, throw, force, pull, begin, need, rocket

Table 5

In Table 5, it can be observed that football quarterbacks and tennis players are distinguishable based on their action patterns. Accordingly, given an unknown person name, if the entity’s associated actions match closer to, for example, the action patterns of known football players, then this person is more likely to refer to a football player than other types of players.

Using this technique, the ERDS can assign entity profiles to one or more predefined categories (roles in this scenario) using inductive learning techniques to automatically construct classifiers based on labeled training data.

The training data is a set of entities that have been manually assigned corresponding categories. Therefore, for each category, there is a set of positive examples, as well as negative examples (entities that don't belong to a particular category). Each profile can then be represented as a feature vector of terms collected from the entity's modifiers and actions. The terms can be weighted in a number of ways, such as:

- binary – a term either appears or does not appear in a profile
- frequency – number of times a term appears in a profile
- other type of weighting scheme

Table 6 below illustrates an example of two feature vectors collected from a segment of text. In this example, the number in each cell indicates the frequency the term (column) occurs in the profile of an entity (row) collected from the text. Table 6 demonstrates that the action profile of an entity named Peyton Manning is more similar to those of known quarterbacks (see example above) than to the profile of Will Smith.

	pass	lead	start	tackle	play	complete	throw
Peyton Manning	2	2	1	0	1	1	3
Will Smith	0	1	1	2	1	0	0

Table 6

Given a set of labeled feature vectors, the ERDS builds a binary classifier (true or false) for each category (e.g., role). The binary classifier can then be directly compared to a generated entity profile to determine whether the entity profile "is one" or "is not one" of a designated category. A number of machine learning algorithms can be used to construct the classifiers, such as Nearest Neighbors, Decision Trees, Support Vector Machines, Maximum Entropy Models, and Rocchio's. In one embodiment, the building of category classifiers is done before new text is processed.

When processing a text segment, the ERDS performs the same feature extraction on each recognized entity to generate feature vectors. Then, the trained classifiers are applied to the extracted feature vectors to determine matches. Corresponding categories are then assigned to each recognized entity

based on the classifier output (*i.e.*, true or false for each category). An entity can be assigned zero, one, or more categories. At the end of this stage (corresponding to the end of block 402 in Figure 4), a set of candidate entities are produced based upon the assigned categories.

5 Determination of Candidate Entities - Ontology/Knowledge Repository Lookup

Referring again to Figure 4, in block 403 the entity recognition and disambiguation process (as performed by the ERDS) maps the entity profiles (of detected entity mentions) to candidate entities (*i.e.*, entries) specified in knowledge bases to assist in the resolution of which precise entity is being referred to (disambiguation). In order to perform this mapping, it is helpful to first understand how the knowledge bases are organized and how corresponding knowledge base entries are discovered. Disambiguation is described further below.

Knowledge about entities can come from many different sources (*e.g.*, gazetteers for geographic location names, databases of person names, census data, etc.) and appear in very different formats. The size of the knowledge bases can be very large. For example, the gazetteer produced by the National Geospatial-Intelligence Agency (NGA) alone contains information about millions of geographic locations that are outside of the United States. Therefore, it is not practical to access the knowledge sources directly, nor it is efficient to search for entities from the sources. Accordingly, the InFact®/Evri search technology has developed and uses an ontology as an repository to integrate such information from multiple, heterogeneous sources. The ontology repository provides a unified interface for finding entities and corresponding information that is derived from multiple sources and stored in the repository. For many purposes, the entity locating or lookup process needs to be real-time, substantially real time, or “near real-time.” Other types of knowledge repositories for combining the information from the multiple sources and for providing a consistent interface to potentially heterogeneous information may be used as well, including other types of data structures other than databases, inverted indices, etc. For the examples described herein, inverted ontology indices have yielded efficient results.

The ontology repository contains a list of entities and information about the entities. Each ontology entry may contain information such as one or more of the following information data:

- Name and synonyms of the entity
- List of types assigned to the entity (e.g., City, Country, Politician)
- Unique identifier and link to original knowledge sources
- 5 • List of other entities that are related to the entity
- Domain-specific, semantic relations (e.g., part-of relations for geographic locations, part-of (Seattle, Washington))
- 10 • Placeholder for storing more detailed description of the entity. The description could be represented as natural language text or as a list of phrases.
- Placeholder for name-value pairs of other attributes (e.g., coordinates of geographic location)

An example of the ontology entry for the state of Washington is shown in Table 7 below.

15

Attribute Name	Value
Name	Washington, State of Washington, WA, Wash., Evergreen State
Type	Province
Part-of-country	United States
Knowledge source and ID	USGS-1779804
Coordinate	47.5, -120.5

Table 7

Ontology Indexing and Lookup

In an example embodiment, the developed ontology repository consists of two components:

- 20 (1) An off-line indexing component that creates inverted indices of the ontology entries in order to enable efficient searching. The indexing is scalable to large size broad-coverage data/knowledge repositories (such as a knowledge base).

(2) An on-line search component that, given an entity name as query, as well as certain attributes of the entity, returns all matching ontology entries.

To be practicable for use with the entity recognition and disambiguation system/process, the on-line search (ontology lookup) preferably has very low latency, so that processing each document can be fast (which could involve many lookups). An inverted index is an index structure storing a mapping from words to their locations in a set of documents. The goal is to optimize the speed of the query: to quickly find the documents where word "X" occurs. Inverted index data structures are often used in document retrieval systems (such as Web search engines). For use with the ERDS, each ontology entry is considered a document, and the entity names and attributes of the ontology entry are treated as terms (*i.e.*, keywords) in the document. Thus, when keywords (for example, from the entity profiles) are specified to the ontology search component, the search component retrieves all "documents" (*i.e.*, ontology entries) with matching terms (entity names and attributes). Using the inverted indexing structure, the system can scale to large ontology sizes, perform rapid lookups across potentially thousands, hundreds of thousands or millions of entities matching a particular criterion.

During the ontology lookup, a search usually specifies an entity name (*e.g.*, "Will Smith"), as well as certain attribute constraints associated with the entity (*e.g.*, football player). The search can be handled using at least two basic approaches:

- A query can be generated to contain only the entity name. After retrieving the list of ontology entries, some post filtering can be applied to the retrieved list by going through the returned entries and selecting the ones matching the specified attribute constraints.
- In addition to entity name, the query can be generated to include partial or all of the attribute constraints. Accordingly, the attribute constraints are resolved during the search. For a common entity name (*e.g.*, person name "John Smith") that might return hundreds, or even thousands of matches, the second approach allows rapidly retrieval of a small list of candidates to work with.

Combinations of these approaches are also possible.

Collecting Related Entities

For each ontology entry, the ontology may store related entities that can be leveraged as additional context clues during the entity disambiguation process (described in the next section). The list of related entities can be generated in a number of ways, including at least the following:

- constructed manually;
- derived from structured knowledge bases; for example, for movie actor "Will Smith", related entities could include the titles of major movies he acted in, other actors he co-starred with, etc.; or
- automatically collected from a set of text documents.

A number of techniques are available to automatically determine related entities from a large corpus of text documents. One technique involves the periodic execution of IQL (InFact® Query Language, now referred to as EQL for Evri Query Language) based "tip" queries which result in related entities based on a relevance ranking of clause level entity binding. As additional unstructured text arrives into the system, related entities may change. The InFact®/Evri tip system dynamically evolves the related entities as the text changes. Example tip systems are explained in more depth in U.S. Patent Publication No. 2007/0156669A1, and U.S. Patent Application No. 12/049,184, filed March 14, 2008.

In addition to clause level entity relations, another automated technique leverages word or entity co-occurrence, via techniques such as latent semantic regression ("LSR") across the incoming unstructured text to provide additional context hints such as terms or entities which tend to occur near the entity of interest. See, for example, U.S. Patent Nos. 6,510,406, issued January 21, 2003; 6,757,646, issued June 29, 2004; 7,051,017, issued May 23, 2006; and 7,269,598, issued September 11, 2007.

Disambiguation – Identify Entities

Once the possible candidate ontology entities (entries in the ontology repository or other knowledge repository) have been retrieved, which could be many (since often entities share the same name), the next step is to determine which of the entities is actually being referred to by the entity mention in the text. The disambiguation process matches the characteristics of the entity described in the text against the attributes of the entities retrieved from the ontology to make this determination. In one example embodiment, the disambiguation process determines a ranking of the candidates, by computing ranking scores based on a

combination of several factors, optionally revisiting the ranking as additional information is gleaned.

Figure 8 is an example flow diagram of an example disambiguation process of an example embodiment. In block 801, the process assigns a default ranking score to each candidate entity based on its importance (see subsection on Default Preferences). Then, in blocks 802-806, the disambiguation process examines each recognized entity and attempts to disambiguate it – to choose the best matching candidate entity to more precisely identify the entity in the text segment. In some embodiments, it is not desirable to choose a matching candidate in a first pass, thus the entity in the text segment may be returned to the list to be processed. The list is thus processed unless no entities remain unresolved or until the results do not change (no additional information is learned). Alternatively, in some embodiments, multiple passes of the disambiguation process are performed on some or all of the entities in the text segment until no change is detected. This embodiment is reflected in Figure 8 in blocks 807-808.

More specifically, in block 802, the process determines whether there are additional entities from the text to be disambiguated, and, if so, continues with block 803, else continues with block 807. In block 803, the disambiguation process gets the next entity from the text segment to process. In block 804, the disambiguation process matches the characteristics of the entity being processed to the matching candidate entities (one or more entries in the ontology) and finds the best possible match. As further described below, this matching process may be solved as a classification problem. In embodiments that delay processing of unresolved entities, in block 805, the process returns the best matching candidate or returns the text segment entity back to the list to be resolved later.

In block 806, the disambiguation process updates the ranking of candidate entities based upon information learned in the process. For example, the ranking may be updated as a result of matching the characteristics of the entity described in the text against the attributes of the candidate entities retrieved from the ontology. As the disambiguation process examines each entity in the document, more information is gained about other entities in the document. Therefore, in some embodiments, it makes sense to revisit the previously resolved entities to check whether to update the previous decision in light of new information. Examples are provided below. The disambiguation process then returns to the beginning of the loop in block 802.

If there were no more entities in the text to disambiguate, then in block 802, the process continues with block 807. Here, in embodiments that

support a multi-pass approach, the process determines whether more passes are needed and if so returns to the beginning of the loop in block 802. If not, the process continues in block 808 to return the best matching candidates for all entities, and ends. These best matching candidate entities are considered the
5 identified entities in Figure 4.

The specifics of each of these blocks are described in more detail in the following subsections.

Default Preferences

The default ranking (preferences) of matching candidates (block 801)
10 is based on each candidate's popularity or importance. For example, a city with a large population or a capital (*e.g.*, Paris, France) is typically ranked higher than other cities, unless there is evidence indicating otherwise. Similarly, a celebrity is more likely to match a person name than a less-known person with an identical name.

15 The importance of an entity can be determined using one or both of two approaches:

(1) Using a statistical approach, which considers how frequently an entity is mentioned in a large corpus of text (*e.g.*, news articles, web blogs, etc.), as well as how much information exists about the entity in the knowledge
20 sources. For example, the more famous a person, the more likely that detailed information is available about the person in the knowledge repository.

(2) Using a knowledge-based approach, which considers certain domain-specific factors, such as:

- Type or role of an entity; *e.g.*, an NFL football player is a
25 more likely match than a high school athlete; Germany the country is a more likely match than a city named Germany (could be found in Georgia, Indiana, and Texas).
- Numeric values associated with an entity; *e.g.*, population of cities, number of movies an actor acted in.
- 30 • "Top N" factor - whether or not an entity appears in a top N list (*e.g.*, the top 100 movie actors, world's largest cities, etc.) which are readily available for many domains.
- Number of inbound links; *e.g.*, in Wikipedia, total number of pages that contain hyperlinks pointing to a given entity
35 usually indicates how important or popular the entity is.

- Whether or not an entity is referred to by its common or default name; e.g., the name Bob Dylan is more likely referring to the musician than his birth name Robert Zimmerman.
- 5
- Whether an entity is current (e.g., a person is alive or a historic figure.)

Disambiguation as Classification Problem

In block 804, the disambiguation process examines the characteristics of the entity described in the text against the attributes of the candidate entities retrieved from the ontology to determine the best matching candidate entity. One approach to the disambiguation process is to use classification algorithms to distinguish between various candidate entities.

In a classification problem, each data instance is assigned a class. A classification model is constructed based on training data, and this model can be used to assign a class to an unseen data instance. The classification model is constructed based on features (properties) of data instances. For disambiguation purposes, each entity can be assigned to either a RELEVANT or a NON-RELEVANT class.

The classification model is built based on manually extracted ground truth (*i.e.*, facts that are known). For example, to construct a classification model, the text in a given document is processed and candidate entities are presented to a human annotator. The annotator then assigns a class to each candidate entity ("RELEVANT" or "NON-RELEVANT"). This classification, together with feature vectors that represent the training data (candidate entries) are stored in a machine readable format. Based on this stored data, a classification model is built using statistical and machine learning techniques.

Once the classification model is constructed for disambiguation purposes, the disambiguation process can then classify newly observed data instances, such as to distinguish between the candidate entities. Newly observed data instances (*i.e.*, the feature vectors generated therefore) can be compared to the stored data using the learning algorithms that implement the classification model. Those that compare more favorably will emerge as better candidates.

More specifically, to perform such disambiguation, as described above, the disambiguation process extracts candidate entities based on string matches to entities in the ontology repository. For example, the string "Will Smith" may match ontology entries for an actor/singer, a football player, a British

comedian, a cricket player, or it may represent none of these people. Further, the matching process may apply anaphora resolution to resolve acronyms, first and last names, or abbreviations to full names of entities.

For example, "Will Smith" may be later referred to as "Smith" or as
5 "Will". Without anaphora resolution, one would have to disambiguate "Smith" between movies with this title, music albums, or multiple people with this same last name. When anaphora resolution is applied, the number of possible candidates may be reduced based on a full name of an entity.

In addition to reducing the number of candidates, the disambiguation
10 process may be enriched through discovered aliases. For example, when the process discovers that CIA is an alias for Culinary Institute of America, it is able to disambiguate CIA to the Culinary Institute and not to the Central Intelligence Agency.

As an example, consider the following text:

15 "Will Smith was born in Queens, NY. He is an American football defensive end. Smith currently plays for the New Orleans Saints of the NFL."

Through coreferencing and anaphora resolution, the recognition and
20 disambiguation process resolves that "Smith" in the third sentence refers to "Will Smith", and does not further consider candidate entities with the name "Smith" and no first name. Through natural language processing, the entity recognition process establishes that "Will Smith" is a person, that "Will Smith" is a male, and that he is football defensive end. It also determines that the string "Will Smith" is a noun and
25 is a subject in the sentence.

As described in the previous section, each entity in a knowledge repository is associated with a particular role or entity type. Examples of roles include being an actor, an album, a band, a city, a football player, a movie, etc. Each entity may also have other associated properties, which can be stored in the
30 knowledge repository for later retrieval during the classification process. For example, "Will Smith", the football player, may have associated properties such as the name of the team(s) that he plays for and his birthplace. Meanwhile, "Will Smith", the movie actor, may have associated properties such as the names of the movies he starred in. Accordingly, such attributes associated with a subject can
35 help aid in future identification and in the classification of which "Will Smith" has been located, tagged, or otherwise referred to.

For each of the candidate entities matched from the knowledge repository, the disambiguation process constructs a feature vector using:

- 5
 - Part of speech of a phrase (e.g., noun, adjective, etc.);
 - Grammatical role of a phrase (e.g., subject, object, prepositional complement, etc.);
- 10
 - General entity types (e.g., person, location, organization, date, number, etc.) discovered through Natural Language Processing;
 - Capitalization of the phrase (more specifically, with regard to the capitalization of a sentence containing the phrase);
 - Presence of quotes and/or brackets around the entity candidate;
 - Number of words in a phrase;
- 15
 - Number of entities in a document potentially related to a candidate entity (for the example above, the entities "New Orleans Saints", "NFL", "Queens", and "NY" are potentially related to "Will Smith."); (The knowledge repository may store a list of related entities for each ontology entry as described above.);
- 20
 - Role of entity (e.g., actor, director, musical artist, politician, movie, album, etc.);
 - Presence of phrase in a standard English lexicon;
 - Default importance of an entity;
- 25
 - Noun modifiers, appositive modifiers, prepositional modifiers of the phrase (e.g., in the sentence: Movie "Smith" was released in 1987, Smith is modified by movie);
 - IDF (inverse document frequency) of a phrase;
 - presence of special characters (comma, dash) in the phrase;
- 30
 - If an entity is a person, using information such as: a) how common is the first and last name, and b) whether the person is alive;

- 5 • If an entity is used in a sentence in grammatical conjunction, using information about the distribution of candidate entities in this conjunction (e.g., are most of the candidates of the same type as the entity? are they all actors? etc.);
- Words in the phrase (the first and last word of a phrase is often indicative of entity type, e.g., first word "Honda" may indicate a vehicle and last word "railway" may indicate a transportation company); and
- 10 • Coreferencing information about entities (e.g., if both "I am Legend" and "Legend" are mentioned in a text as movies, the "Legend" more likely refers to "I am Legend" than to a different movie also titled "Legend.")

15 An example of feature vectors for possible "Will Smith" candidates is presented in the Table 8, below:

EntityID	Role	Number of Relevant Entities	Standard Type	Default Importance	Relevant
123112	actor	0	Person	5	No
1232134	football player	3	Person	2	Yes
123788	comedian	1	Person	2	No

Table 8

20 Such feature vectors are then analyzed using the machine learning and statistical learning capabilities of the previously constructed classification model, to classify each feature vector to determine the more likely candidates. For example, in Table 8, the "Relevant" column is populated after the feature vector is run through the classification model. Thereafter, one or more selected feature vectors also can be stored in machine readable form and used to further inform the

25 model.

Note that in some cases, a large number of potential candidates may exist for an entity. In such cases, it may be beneficial to apply a multi-level approach to classification. According to one approach, the disambiguation process may filter possible candidate entities based on relevant entities that exist
5 in the text and then apply the classification model to classify only entities that pass the filter step. Also, different data structures, such as inverted indices, may be used to speed-up the filtering process. In addition, only a portion of, or additional characteristics may be generated as part of the feature vector to compare with the recognized entity under examination.

10 Multi-Pass Updating Process

As the disambiguation process examines the various entities in the document and performs disambiguation, it collects more evidence (especially regarding related entities), which can be used to modify the initial disambiguation decisions or to complete unresolved entity disambiguations. For example, given
15 the text "Hawks dominate 49ers in San Francisco," the term "Hawks" is ambiguous. It could refer to the NFL team Seattle Seahawks, or the NBA basketball team Atlanta Hawks, or perhaps many others things. Unable to resolve "Hawks," the disambiguation process may choose to leave it as is and move on to other entities. Only after it processes and resolves other names in the context
20 (e.g., observes that San Francisco 49ers is an NFL football team), it has sufficient evidence that the context relates to NFL football. After that determination, the disambiguation process can return to the disambiguation of "Hawks", and assert higher confidence to the candidate "Seattle Seahawks." One method to accomplish such latent resolution is by maintained a list of unresolved entities.
25 Another method is to perform multiple passes of the resolution process from the beginning. Other approaches are also usable.

In one example embodiment the disambiguation process adopts a multi-pass process that iteratively updates the scores assigned to each candidate. During each iteration, it examines the names in a document sequentially and
30 performs disambiguation based on the contextual evidence collected thus far. The process continues until there is no new evidence produced or until the number of iterations reaches some determined or predetermined limit. To detect this condition, the process checks whether or not the disambiguation of any entities within the context have changed. If so, any new information is included to the
35 subsequent disambiguation of other entities. If not, the disambiguation process completes.

Updating the disambiguation results is based on assumptions such as the existence of parallelism and focus in the surrounding context. Parallelism means similar entities within the same context are more likely to have the same type. For example, when "Ali" is mentioned together with a list of known movie titles, one could be fairly confident that it refers to the name of a movie in this context instead of a person. Geographic focus or a document's main subject can be computed through a majority vote from entities in the context. For example, suppose a text segment mentions several place names, such as Salem, Portland, and Vancouver, where each name is fairly ambiguous. Table 9, below, shows the top and second candidates resolved after the first pass:

Name	Top candidate	Second candidate
Salem	Salem, Massachusetts	<u>Salem, Oregon</u>
Portland	<u>Portland, Oregon</u>	Portland, Maine
Vancouver	Vancouver, British Columbia	<u>Vancouver, Washington</u>

Table 9

Based on the recognition that all three names have candidates as cities near the Portland, Oregon area (Vancouver, WA lies just north of Portland, OR), the updating process could boost the scores for those candidates. As a result, the disambiguation process would select the second candidate for Salem and Vancouver.

The final result of this process (block 808) is a list of entities (*e.g.*, phrases) in the text linked to their most appropriate ontology entries. A link in this sense may be any form of reference or identification of the corresponding ontology entry.

Ranking/Grouping Identified Entities

After the entity recognition and disambiguation process, many entities could be identified from a given document. Instead of simply listing all the entities, an example embodiment of the recognition and disambiguation system may organize or summarize the entities within a document (see block 405 of Figure 4) by, for example:

- Ranking – sorting the set of entities based on a certain order (e.g., importance within the document), then presenting the top ranked entities;
- Grouping – grouping the set of entities into subsets that share certain properties (e.g., place names of same country or region)

Ranking

Given a list of entities tagged in a document, this process ranks the entities based on their relevance to the main subject or focus of the document.

Features extracted for such a ranking process may include one or more of the following features:

- Frequency of the entity occurring in the document;
- Location of the entity in the document, e.g., in title, first paragraph, caption, etc.;
- Uniqueness of the entity in the document versus in a collection or corpus. One way to measure the uniqueness is to compute inverse document frequency (IDF). For example, it is possible to determine that entities such as "United States" or "Associated Press" are very common, and therefore, should be ranked lower;
- Relevance to main focus or theme of the surrounding text (e.g., geographic focus - "this document is about Iraq", document subject - politics, entertainment, etc), which can be automatically extracted from the document or provided as document metadata attributes;
- Confidence score computed from previous steps; or
- Predefined preferences.

A ranking score can be computed based on a weighted combination of one or more of the above factors. In some embodiments, the ERDS builds a regression model to automatically determine the values of the weights that should be assigned to each factor. A regression method models the relationship of a response variable (ranking score in this case) to a set of predicting factors. In preparation, the ERDS first collects a set of training data that consists of a set of documents, where the ranking of the top entities in each document has been

manually verified. Then, using the training data, it automatically estimates parameters (*i.e.*, weights) of the regression model. Then, when presented with a previously unseen document, the ERDS can then apply the regression model to entities found in the document and predict each entity's ranking score.

5 Grouping

The recognition and disambiguation system can also organize the resulting entities by grouping them based on certain properties associated with the entities. One property for grouping purposes is an entity's type. For example, entities can be grouped into categories such as places, people, and organizations.

10 The groupings can be further organized into an hierarchy. For example, place names can be further grouped into cities, countries, regions, etc. Accordingly, given a document, the entities identified in the document can be viewed as groups in different granularities of the hierarchy.

Another property for grouping is the "part-of" (or "member-of") relation. For example, given an article describing football games, the players mentioned in the text can be grouped according to the football teams they belong to. Similarly, for place names, they can be grouped based on the countries or regions they are part of.

Other grouping and/or ranking techniques can similarly be incorporated, and used by the ERDS to present the entities resulting from the disambiguation process.

Figure 9 is an example block diagram of an example computing system that may be used to practice embodiments of a recognition and disambiguation system such as that described herein. Note that a general purpose or a special purpose computing system may be used to implement an

25 RDS. Further, the RDS may be implemented in software, hardware, firmware, or in some combination to achieve the capabilities described herein.

Computing system 900 may comprise one or more server and/or client computing systems and may span distributed locations. In addition, each

30 block shown may represent one or more such blocks as appropriate to a specific embodiment or may be combined with other blocks. Moreover, the various blocks of the RDS 910 may physically reside on one or more machines, which use standard (*e.g.*, TCP/IP) or proprietary interprocess communication mechanisms to communicate with each other.

In the embodiment shown, computer system 900 comprises a

35 computer memory ("memory") 901, a display 902, one or more Central Processing

Units ("CPU") 903, Input/Output devices 904 (e.g., keyboard, mouse, CRT or LCD display, etc.), other computer-readable media 905, and network connections 906. The RDS 910 is shown residing in memory 901. In other embodiments, some portion of the contents, some of, or all of the components of the RDS 910 may be stored on and/or transmitted over the other computer-readable media 905. The components of the RDS 910 preferably execute on one or more CPUs 903 and perform entity identification (recognition and disambiguation), as described herein. Other code or programs 930 and potentially other data repositories, such as data repository 920, also reside in the memory 910, and preferably execute on one or more CPUs 903. Of note, one or more of the components in Figure 9 may not be present in any particular implementation. For example, some embodiments embedded in other software may not provide means for other user input or display.

In a typical embodiment, the RDS 910 includes a linguistic analysis engine 911, a knowledge analysis engine 912, a disambiguation engine 913, an NLP parsing engine or preprocessor 914, an RDS API 917, a data repository (or interface thereto) for storing document NLP and disambiguation data 916, and a knowledge data repository 915, for example, an ontology index, for storing information from a multitude of internal and/or external sources. In at least some embodiments, the NLP parsing engine / preprocessor 914 is provided external to the RDS and is available, potentially, over one or more networks 950. Other and or different modules may be implemented. In addition, the RDS 910 may interact via a network 950 with applications or client code 955 that uses results computed by the RDS 910, one or more client computing systems 960, and/or one or more third-party information provider systems 965, such as purveyors of information used in knowledge data repository 915. Also, of note, the knowledge data 915 may be provided external to the RDS as well, for example in an ontology knowledge base accessible over one or more networks 950.

In an example embodiment, components/modules of the RDS 910 are implemented using standard programming techniques. However, a range of programming languages known in the art may be employed for implementing such example embodiments, including representative implementations of various programming language paradigms, including but not limited to, object-oriented (e.g., Java, C++, C#, Smalltalk), functional (e.g., ML, Lisp, Scheme, etc.), procedural (e.g., C, Pascal, Ada, Modula, etc.), scripting (e.g., Perl, Ruby, Python, JavaScript, VBScript, etc.), declarative (e.g., SQL, Prolog, etc.), etc.

The embodiments described use well-known or proprietary synchronous or asynchronous client-server computing techniques. However, the

various components may be implemented using more monolithic programming techniques as well, for example, as an executable running on a single CPU computer system, or alternately decomposed using a variety of structuring techniques known in the art, including but not limited to, multiprogramming, 5 multithreading, client-server, or peer-to-peer, running on one or more computer systems each having one or more CPUs. Some embodiments are illustrated as executing concurrently and asynchronously and communicating using message passing techniques. Equivalent synchronous embodiments are also supported by a RDS implementation.

10 In addition, programming interfaces to the data stored as part of the RDS 910 (*e.g.*, in the data repositories 915 and 916) can be made available by standard means such as through C, C++, C#, and Java APIs; libraries for accessing files, databases, or other data repositories; through scripting languages such as XML; or through Web servers, FTP servers, or other types of servers 15 providing access to stored data. The data repositories 915 and 916 may be implemented as one or more database systems, file systems, or any other method known in the art for storing such information, or any combination of the above, including implementation using distributed computing techniques.

Also, the example RDS 910 may be implemented in a distributed 20 environment comprising multiple, even heterogeneous, computer systems and networks. For example, in one embodiment, the modules 911-914, and 917, and the data repositories 915 and 916 are all located in physically different computer systems. In another embodiment, various modules of the RDS 910 are hosted each on a separate server machine and may be remotely located from the tables 25 which are stored in the data repositories 915 and 916. Also, one or more of the modules may themselves be distributed, pooled or otherwise grouped, such as for load balancing, reliability or security reasons. Different configurations and locations of programs and data are contemplated for use with techniques of described herein. A variety of distributed computing techniques are appropriate for 30 implementing the components of the illustrated embodiments in a distributed manner including but not limited to TCP/IP sockets, RPC, RMI, HTTP, Web Services (XML-RPC, JAX-RPC, SOAP, etc.). Other variations are possible. Also, other functionality could be provided by each component/module, or existing functionality could be distributed amongst the components/modules in different 35 ways, yet still achieve the functions of a RDS.

Furthermore, in some embodiments, some or all of the components of the RDS may be implemented or provided in other manners, such as at least

partially in firmware and/or hardware, including, but not limited to, one or more application-specific integrated circuits (ASICs), standard integrated circuits, controllers (e.g., by executing appropriate instructions, and including microcontrollers and/or embedded controllers), field-programmable gate arrays (FPGAs), complex programmable logic devices (CPLDs), etc. Some or all of the system components and/or data structures may also be stored as contents (e.g., as executable or other machine-readable software instructions or structured data) on a computer-readable medium (e.g., as a hard disk; a memory; a computer network or cellular wireless network or other data transmission medium; or a portable media article to be read by an appropriate drive or via an appropriate connection, such as a DVD or flash memory device) so as to enable or configure the computer-readable medium and/or one or more associated computing systems or devices to execute or otherwise use or provide the contents to perform at least some of the described techniques. Some or all of the system components and data structures may also be transmitted as contents of generated data signals (e.g., by being encoded as part of a carrier wave or otherwise included as part of an analog or digital propagated signal) on a variety of computer-readable transmission mediums, including wireless-based and wired/cable-based mediums, and may take a variety of forms (e.g., as part of a single or multiplexed analog signal, or as multiple discrete digital packets or frames). Such computer program products may also take other forms in other embodiments. Accordingly, embodiments of the present disclosure may be practiced with other computer system configurations.

All of the above U.S. patents, U.S. patent application publications, U.S. patent applications, foreign patents, foreign patent applications and non-patent publications referred to in this specification and/or listed in the Application Data Sheet, including but not limited to U.S. Provisional Patent Application No. 60/980,747, entitled "NLP-BASED ENTITY RECOGNITION AND DISAMBIGUATION," filed October 17, 2007, are incorporated herein by reference, in their entirety.

From the foregoing it will be appreciated that, although specific embodiments have been described herein for purposes of illustration, various modifications may be made without deviating from the spirit and scope of this disclosure. For example, the methods, techniques, and systems for entity recognition and disambiguation are applicable to other architectures other than an InFact® /Evri architecture or a Web-based architecture. For example, other

systems that are programmed to perform natural language processing can be employed. Also, the methods, techniques, and systems discussed herein are applicable to differing query languages, protocols, communication media (optical, wireless, cable, etc.) and devices (such as wireless handsets, electronic
5 organizers, personal digital assistants, portable email machines, game machines, pagers, navigation devices such as GPS receivers, etc.).

CLAIMS

1. A computer-implemented method for identifying one or more entities in an indicated text segment, comprising:
 - processing the indicated text segment to determine a plurality of terms and their associated parts-of-speech tags and grammatical roles;
 - performing linguistic analysis of the processed text segment to determine one or more potential entity names which are referred to in the text segment;
 - generating, for each potential entity name, an entity profile having one or more associated properties that characterize the entity based upon surrounding context;
 - and
 - determining one or more mostly likely entities which are referred to in the text segment by comparing the entity profiles generated for each potential entity name with one or more candidate entities using both linguistic and contextual information.
2. The method of claim 1 wherein the candidate entities are entity entries retrieved from a knowledge repository.
3. The method of claims 1 or 2 wherein the knowledge repository is an ontology knowledge base.
4. The method of at least one of the above claims wherein the determining of the one or more mostly likely entities which are referred to in the text segment by comparing the entity profiles generated for each potential entity name with one or more candidate entities further comprises:
 - searching a knowledge repository for a set of candidate entities that have similar characteristics to one or more of the generated entity profiles;
 - ranking the candidate entities in the set of candidate entities to determine a set of mostly likely entities which are referred to in the text segment; and
 - providing the determined set of mostly likely entities.
5. The method of claim 4 wherein the ranking weights the candidate entities according to contextual information surrounding portions of the text segment that refer to the potential entity names.

6. The method of claims 4 or 5 wherein the ranking weights the candidate entities according to preference information.

7. The method of at least one of claims 4 to 6 wherein the ranking the candidate entities further comprises using a classification model to classify the candidate entities and to order them based upon closest matches.

8. The method of at least one of the above claims wherein the determining the one or more most likely entities which are referred to in the text segment by comparing the entity profiles generated for each potential entity name with one or more candidate entities using both linguistic and contextual information further comprises:

resolving the one or more most likely entities to a single identified entity by performing iterative comparisons reusing entity recognition information gained from a prior comparison until no new entity recognition information is gained.

9. The method of at least one of the above claims wherein each entity profile comprises a feature vector of terms collected from modifiers and/or actions associated with the potential entity name based upon the linguistic analysis of the processed text segment.

10. The method of claim 9 wherein the terms are weighted based upon at least one of a binary weight or a frequency weight.

11. The method of at least one of the above claims, further comprising providing a ranked list of the determined one or more most likely entities which are referred to in the text segment.

12. The method of at least one of the above claims, further comprising using the determined one or more likely entities to inform a relationship search.

13. The method of at least one of the above claims, further comprising: performing the method in response to selection of a user interface component.

14. The method of at least one of the above claims wherein the method is embedded in code that supports a widget presented on a web page.

15. The method of at least one of the above claims, further comprising:
Invoking the method to annotate information on a web page.

16. A computer-readable medium containing contents that, when executed causes a computing system to perform at least one of the methods of the above claims.

17. The computer-readable medium of claim 16, wherein the medium is a memory in a computing system containing the contents.

18. The computer-readable medium of claims 16 or 17 wherein the contents are instructions that, when executed, cause a computer processor in the computing device to perform the method of at least one of claims 1-15.

19. The computer-readable medium of at least one of claims 16 to 17 embedded in a computing system configured to perform indexing and storing of a corpus of documents for searching using natural language processing.

20. An entity recognition and disambiguation system comprising one or more components for performing at least one of the methods of claims 1-15.

21. An entity recognition and disambiguation computing system, comprising;

a memory; and

a recognition and disambiguation module stored in the memory that is configured, when executed, to

receive a text segment for processing;

recognize one or more candidate named entities which are referred to by a detected entity in a received text segment based, at least in part, upon a natural language analysis of the text segment; and

disambiguate the candidate named entities to determine a single named entity to which the detected entity in the received text segment is deemed to refer based upon a combination of linguistic analysis, contextual information gleaned

from surrounding text, and information retrieved from one or more knowledge repositories.

22. The system of claim 21, wherein the module is further configured, when executed, to disambiguate the candidate named entities using information based upon a relationship search.

23. The system of claims 21 or 22 wherein the module is further configured, when executed, to disambiguate the candidate named entities using a classification modeling approach.

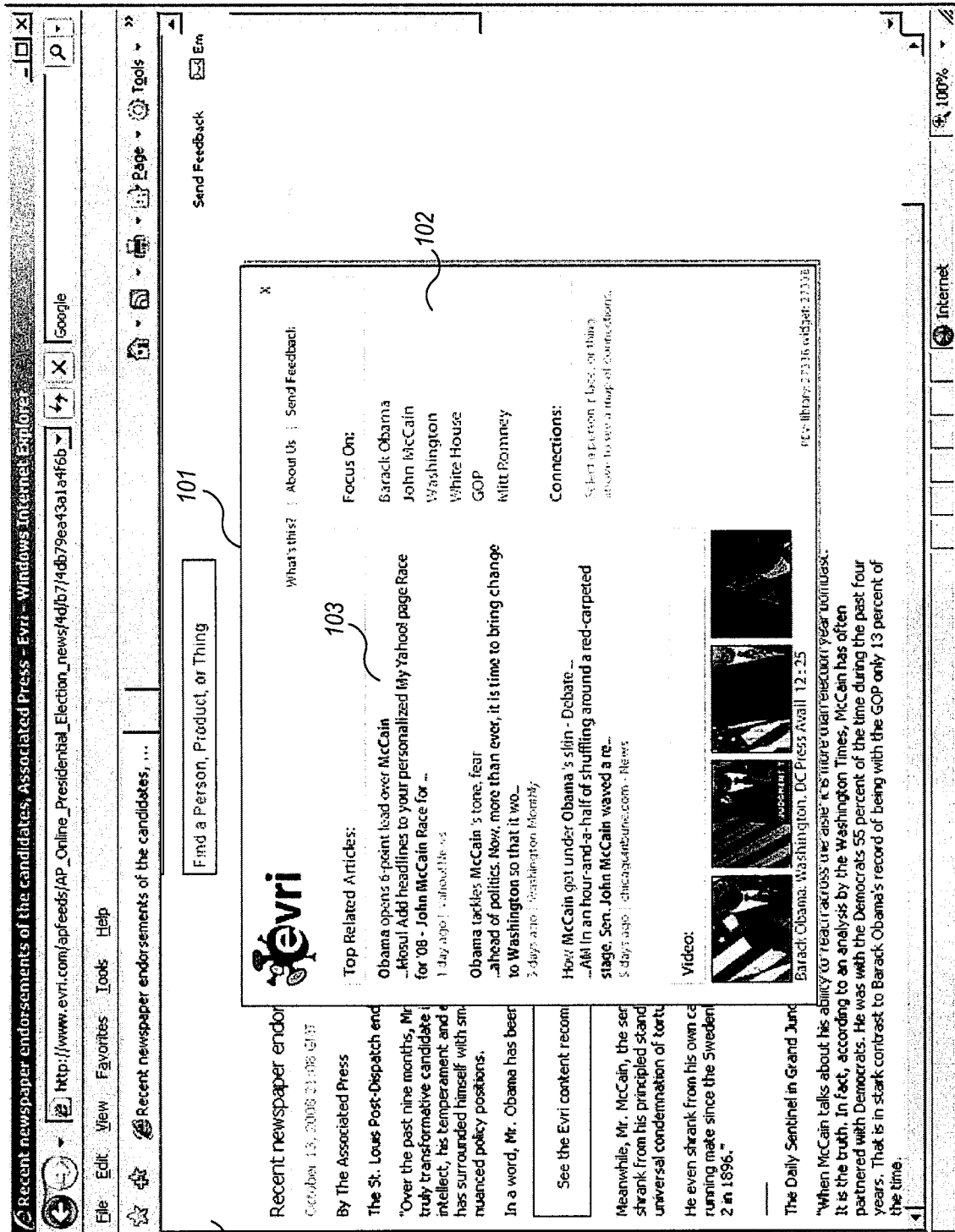
24. A computer-implemented method for presenting information regarding named entities comprising:

receiving an indication of a segment of text;

invoking the recognition and disambiguation module of at least one of claims 20 to 23 to process the indicated text segment to automatically determine one or more named entities referred to in the text segment; and

for each determined one or more named entities, presenting a link to information associated with the named entity.

25. The method of claim 24 wherein the information is based in part upon an ontology entry.



100

110

Fig. 1

201

X

What's this? | About Us | Send Feedback

Top Related Articles: John McCain





McCain vows to fight for new direction for ...
By BETH FOUHY Associated Press Writer VIRGINIA BEACH, Va.
Republican **John McCain** pledged to ...
3 hours ago | Miami Herald | Politics

McCain seeks to revive campaign, reassure ...
... **Double-Standard**. billogden.org Reuters Photo: Republican
presidential nominee Senator John M...
19 hours ago | Yahoo! News

**McCain Asks Voters to Show Respect
more in Politics & Campaign** » Sen. **John McCain** spoke out against
the growing nastiness his...
1 day ago | Wall Street Journal: Pol...

See full profile: **John McCain**

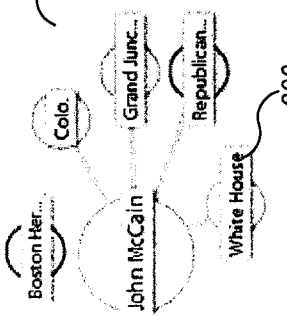
Video:

Focus On:

Barack Obama
John McCain 203
Washington
White House
GOP
Mitt Romney

Connections:



REF: library.23301.206.ec.27331

205

Fig. 2

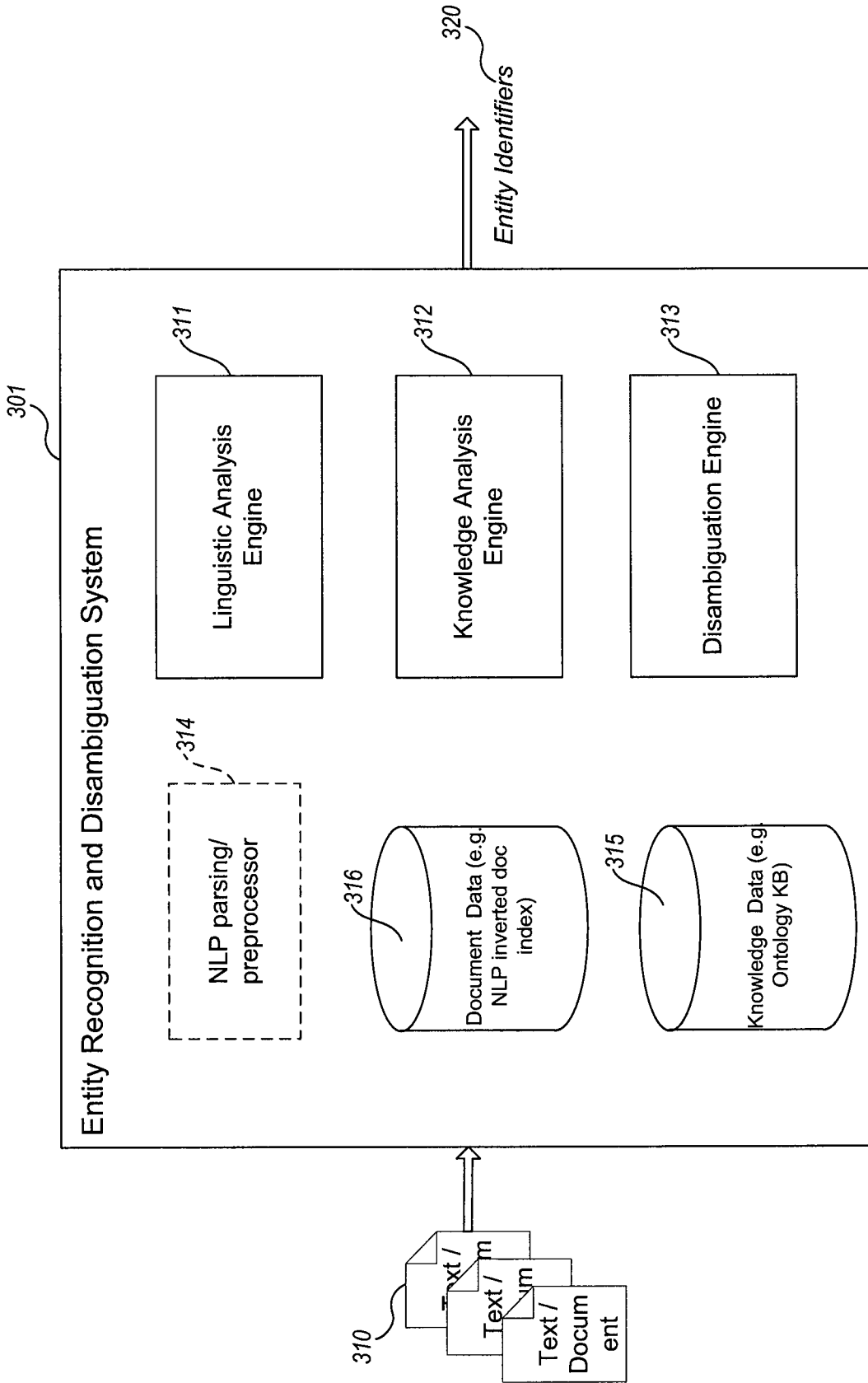


Fig. 3

4/9

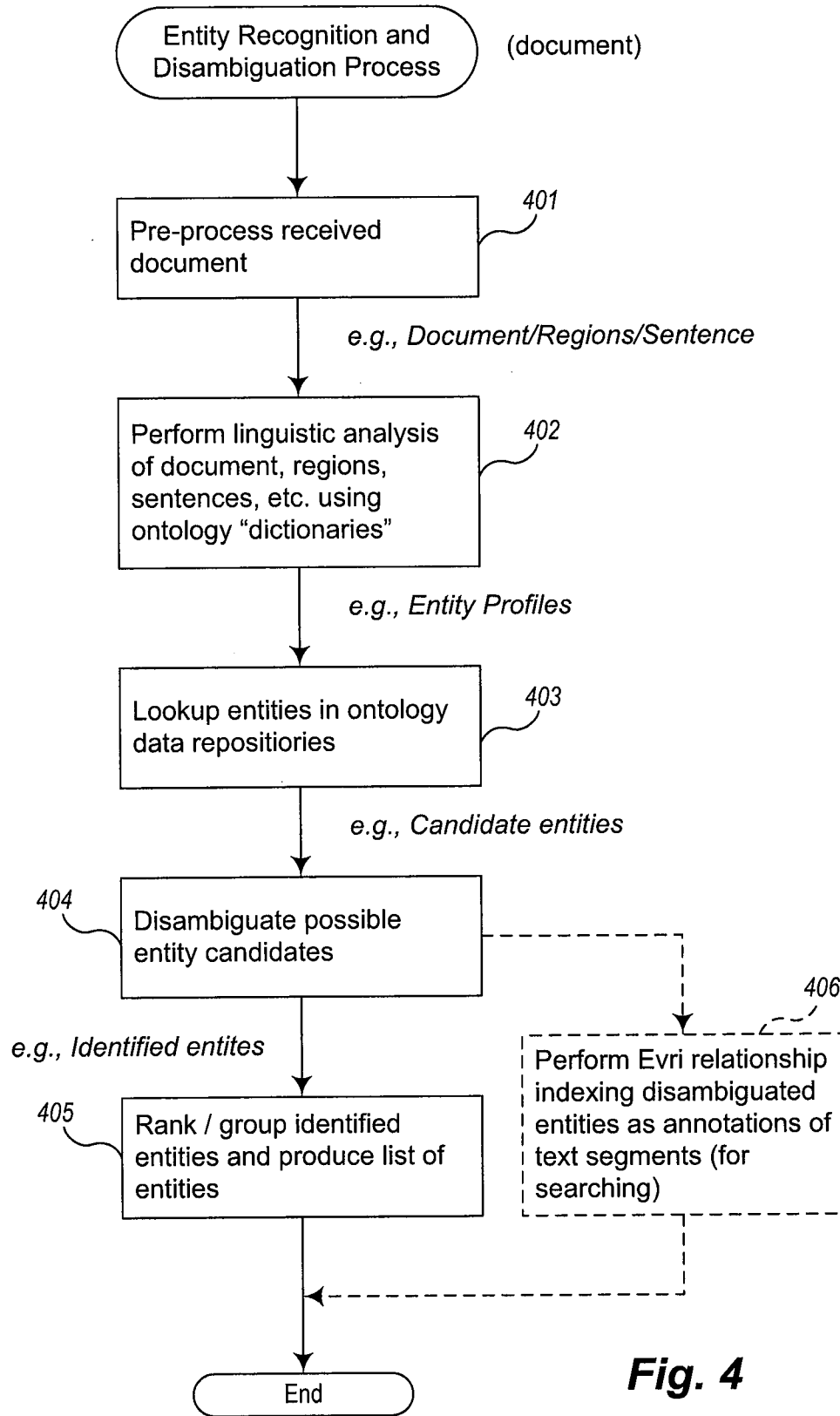


Fig. 4

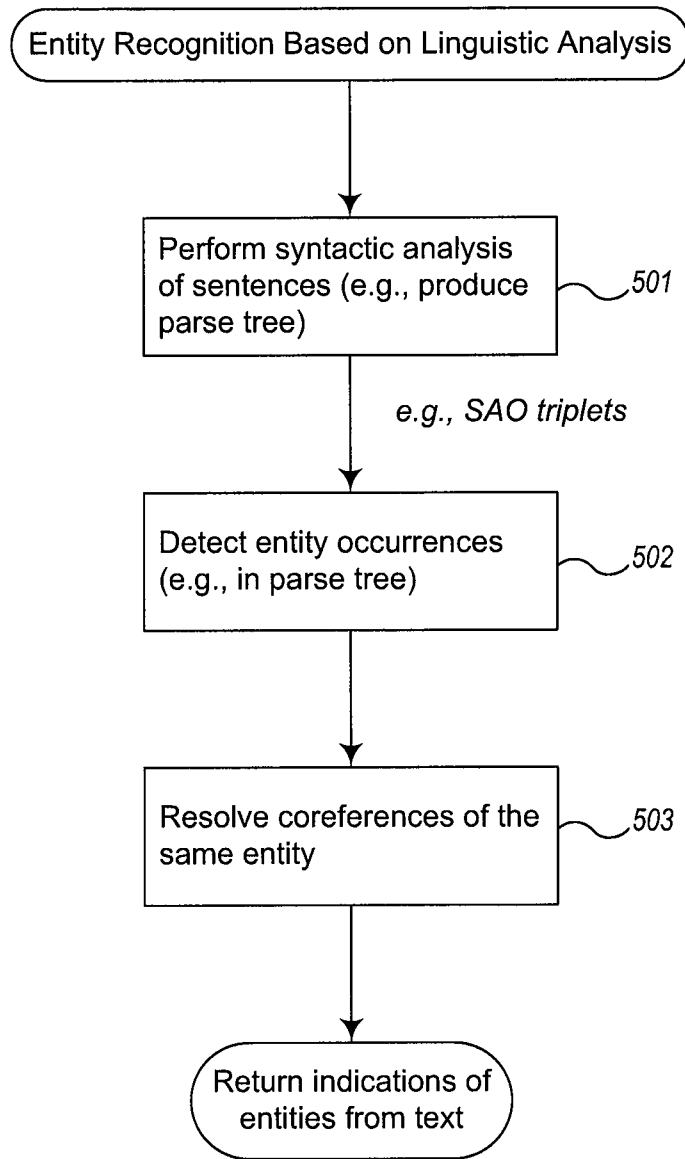


Fig. 5

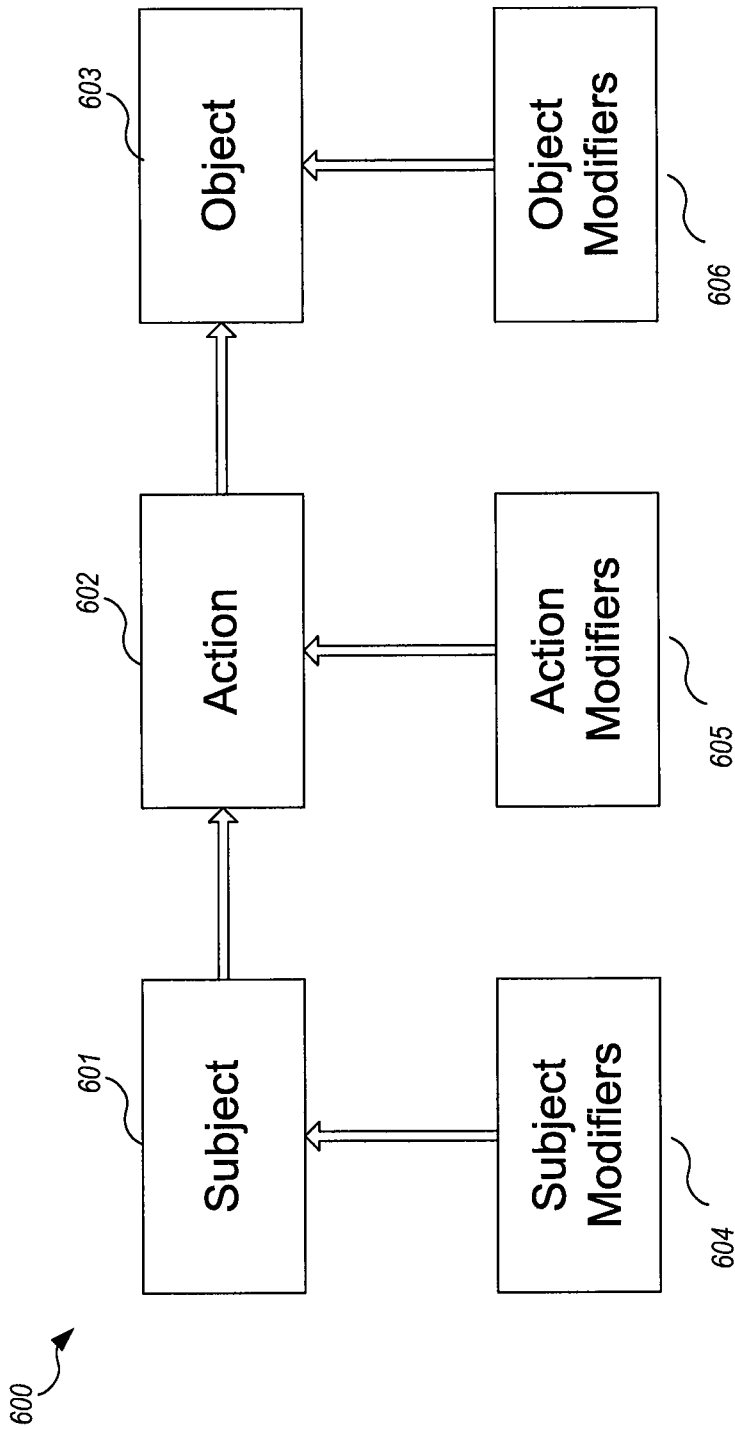


Fig. 6

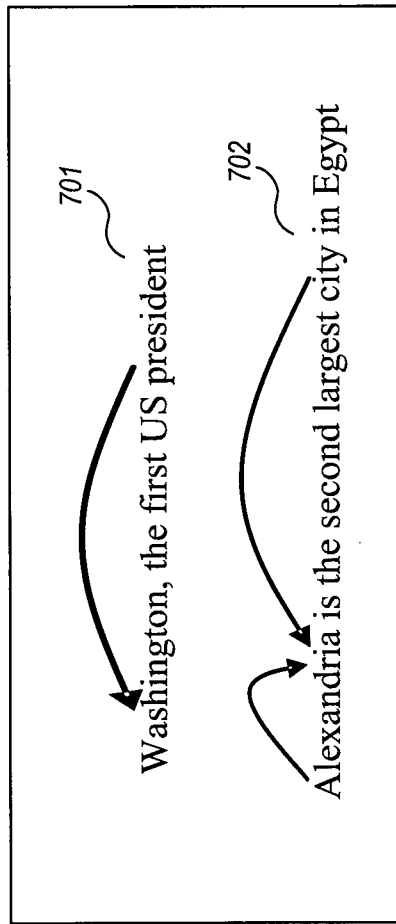


Fig. 7

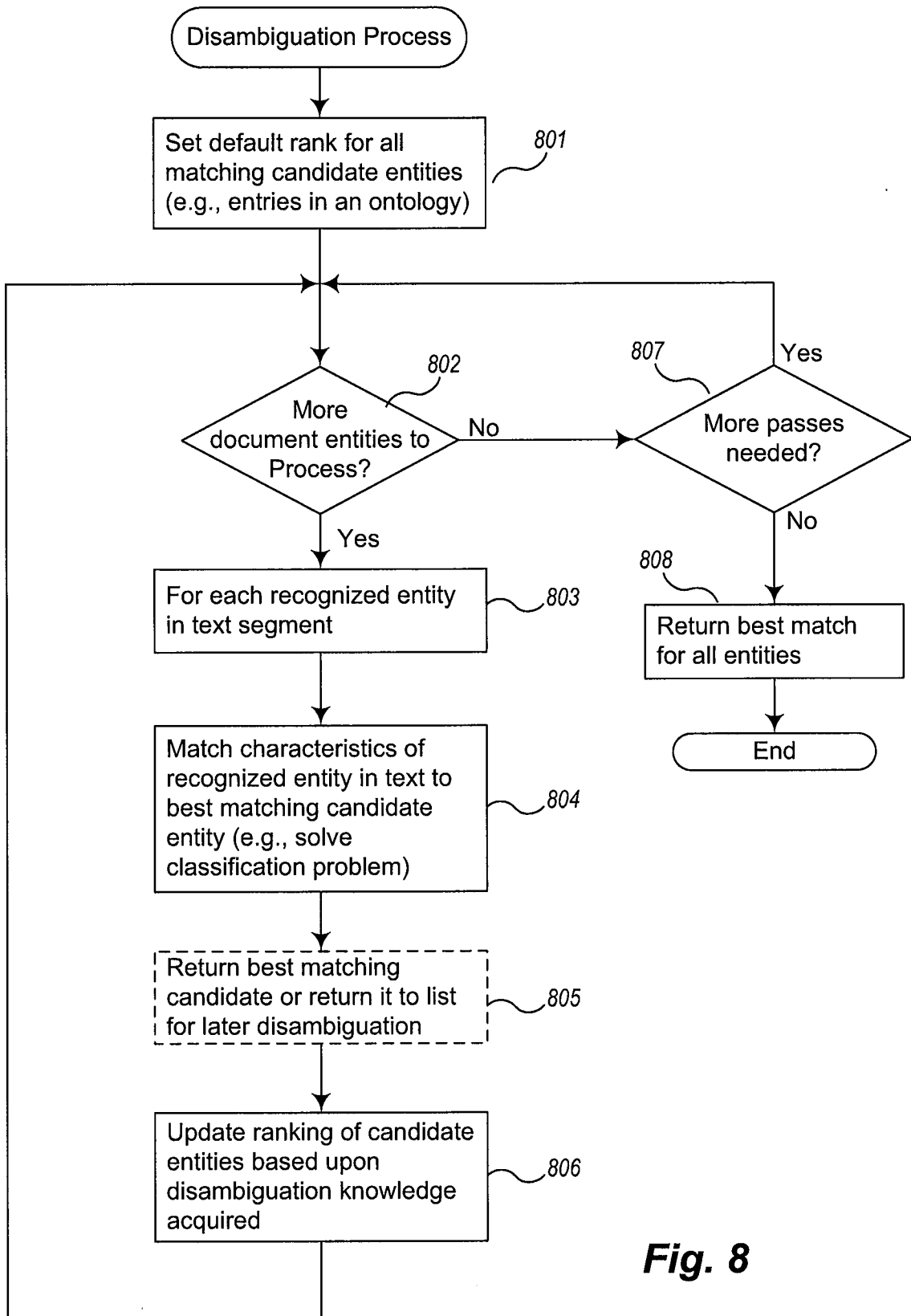


Fig. 8

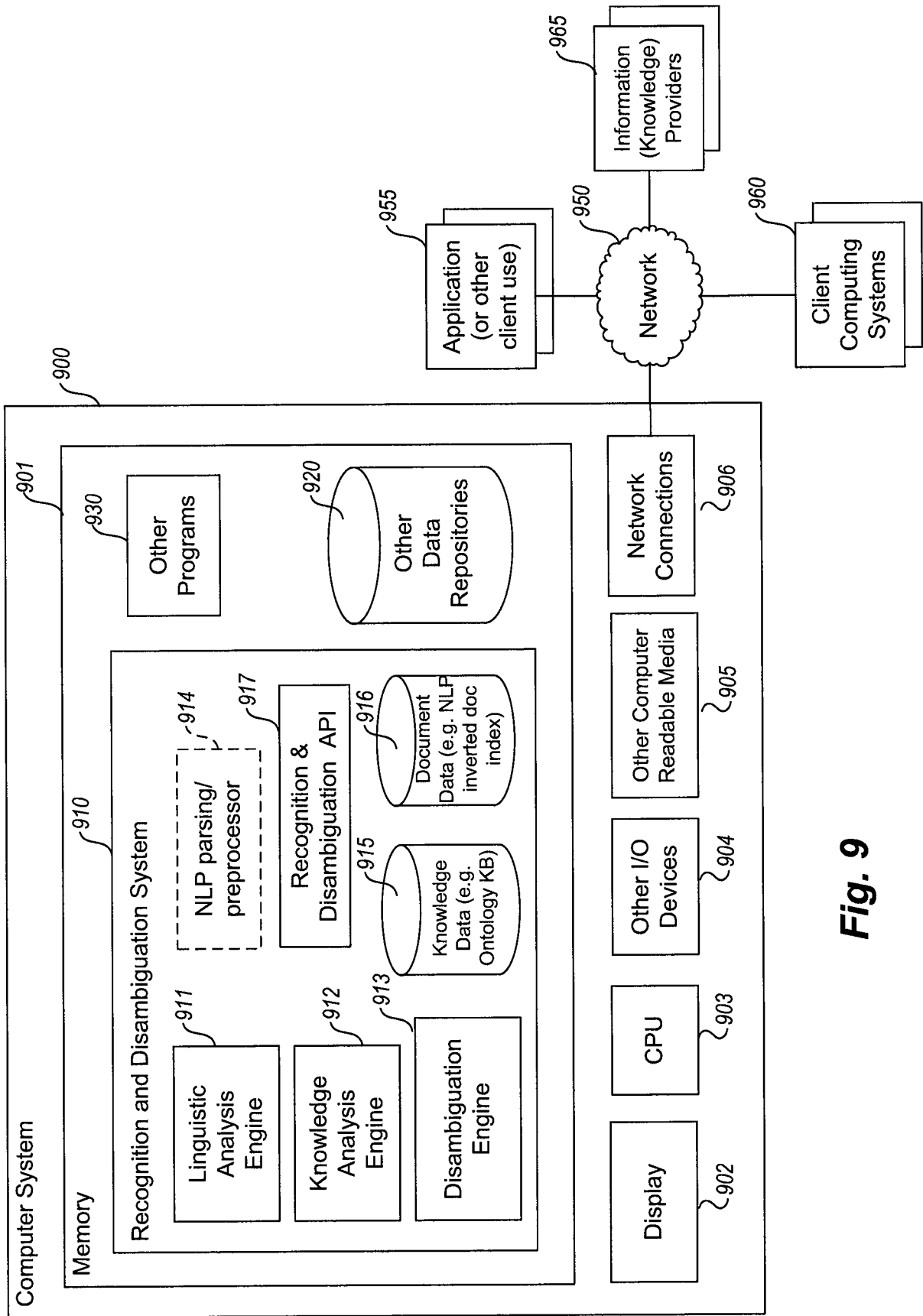


Fig. 9

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US 08/80149

A. CLASSIFICATION OF SUBJECT MATTER

IPC(8) - G06F 17/20 (2008.04)

USPC - 704/1

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)

IPC(8) - G06F 17/20 (2008.04)

USPC - 704/1

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

USPC - 704/7-9; 707/1-3 - search terms below.

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

Google Scholar; WEST (USPT, PGPB, EPAB, JPAB) - extract, entity, identify, identification, distinguish, part of speech, speech classifier, phrase classifier, grammatical, syntactic, semantic, role, function, purpose, character, linguistic, context, surrounding, ontology, metaphysics.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2005/0216443 A1 (MORTON et al.) 29 September 2005 (29.09.2005) - para [0036], [0058]-[0061], [0063], [0065]-[0070], [0172], [0176]-[0177], [0180], [0182]-[0184].	1-3, 21-23
A	US 2005/0197828 A1 (MCCONNELL et al.) 08 September 2005 (08.09.2005).	1-3, 21-23

 Further documents are listed in the continuation of Box C.

* 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

"&" document member of the same patent family

Date of the actual completion of the international search

03 December 2008 (03.12.2008)

Date of mailing of the international search report

17 DEC 2008

Name and mailing address of the ISA/US

Mail Stop PCT, Attn: ISA/US, Commissioner for Patents
P.O. Box 1450, Alexandria, Virginia 22313-1450
Facsimile No. 571-273-3201

Authorized officer:

Lee W. Young

PCT Helpdesk: 571-272-4300
PCT OSP: 571-272-7774

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US 08/80149

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. Claims Nos.:
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:

- 3. Claims Nos.: 4-20 and 24-25
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 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.