10

Automatic Concept Extraction in Semantic Summarization Process

Antonella Carbonaro
Computer Science Department, University of Bologna,
Mura Anteo Zamboni,
Italy

1. Introduction

The Semantic Web offers a generic infrastructure for interchange, integration and creative reuse of structured data, which can help to cross some of the boundaries that Web 2.0 is facing. Currently, Web 2.0 offers poor query possibilities apart from searching by keywords or tags. There has been a great deal of interest in the development of semantic-based systems to facilitate knowledge representation and extraction and content integration [1], [2]. Semantic-based approach to retrieving relevant material can be useful to address issues like trying to determine the type or the quality of the information suggested from a personalized environment. In this context, standard keyword search has a very limited effectiveness. For example, it cannot filter for the type of information, the level of information or the quality of information.

Potentially, one of the biggest application areas of content-based exploration might be personalized searching framework (e.g., [3],[4]). Whereas search engines provide nowadays largely anonymous information, new framework might highlight or recommend web pages related to key concepts. We can consider semantic information representation as an important step towards a wide efficient manipulation and retrieval of information [5], [6], [7]. In the digital library community a flat list of attribute/value pairs is often assumed to be available. In the Semantic Web community, annotations are often assumed to be an instance of an ontology. Through the ontologies the system will express key entities and relationships describing resources in a formal machine-processable representation. An ontology-based knowledge representation could be used for content analysis and object recognition, for reasoning processes and for enabling user-friendly and intelligent multimedia content search and retrieval.

Text summarization has been an interesting and active research area since the 60's. The definition and assumption are that a small portion or several keywords of the original long document can represent the whole informatively and/or indicatively. Reading or processing this shorter version of the document would save time and other resources [8]. This property is especially true and urgently needed at present due to the vast availability of information. Concept-based approach to represent dynamic and unstructured information can be useful to address issues like trying to determine the key concepts and to summarize the information exchanged within a personalized environment.

In this context, a concept is represented with a Wikipedia article. With millions of articles and thousands of contributors, this online repository of knowledge is the largest and fastest growing encyclopedia in existence.

The problem described above can then be divided into three steps:

- Mapping of a series of terms with the most appropriate Wikipedia article (disambiguation).
- Assigning a score for each item identified on the basis of its importance in the given context.
- Extraction of n items with the highest score.

Text summarization can be applied to many fields: from information retrieval to text mining processes and text display. Also in personalized searching framework text summarization could be very useful.

The chapter is organized as follows: the next Section introduces personalized searching framework as one of the possible application areas of automatic concept extraction systems. Section three describes the summarization process, providing details on system architecture, used methodology and tools. Section four provides an overview about document summarization approaches that have been recently developed. Section five summarizes a number of real-world applications which might benefit from WSD. Section six introduces Wikipedia and WordNet as used in our project. Section seven describes the logical structure of the project, describing software components and databases. Finally, Section eight provides some considerations on case study and experimental results.

2. Personalized searching framework

In personalized searching frameworks, standard keyword search is of very limited effectiveness. For example, it does not allow users and the system to search, handle or read concepts of interest, and it doesn't consider synonymy and hyponymy that could reveal hidden similarities potentially leading to better retrieval. The advantages of a concept-based document and user representations can be summarized as follows: (i) ambiguous terms inside a resource are disambiguated, allowing their correct interpretation and, consequently, a better precision in the user model construction (e.g., if a user is interested in computer science resources, a document containing the word 'bank' as it is meant in the financial context could not be relevant); (ii) synonymous words belonging to the same meaning can contribute to the resource model definition (for example, both 'mouse' and 'display' brings evidences for computer science documents, improving the coverage of the document retrieval); (iii) synonymous words belonging to the same meaning can contribute to the user model matching, which is required in recommendation process (for example, if two users have the same interests, but these are expressed using different terms, they will be considered overlapping); (iv) finally, classification, recommendation and sharing phases take advantage of the word senses in order to classify, retrieve and suggest documents with high semantic relevance with respect to the user and resource models.

For example, the system could support Computer Science last-year students during their activities in courseware like Bio Computing, Internet Programming or Machine Learning. In fact, for these kinds of courses it is necessary an active involvement of the student in the

acquisition of the didactical material that should integrate the lecture notes specified and released by the teacher. Basically, the level of integration depends both on the student's prior knowledge in that particular subject and on the comprehension level he wants to acquire. Furthermore, for the mentioned courses, it is continuously necessary to update the acquired knowledge by integrating recent information available from any remote digital library.

3. Inside summarization

Summarization is a widely researched problem. As a result, researchers have reported a rich collection of approaches for automatic document summarization to enhance those provided manually by readers or authors as a result of intellectual interpretation. One approach is to provide summary creation based on a natural language generation (as investigated for instance in the DUC and TREC conferences); a different one is based on a sentence selection from the text to be summarized, but the most simple process is to select a reasonable short list of words among the most frequent and/or the most characteristic words from those found in the text to be summarized. So, rather than a coherent text the summary is a simple set of items.

From a technical point of view, the different approaches available in the literature can be considered as follows. The first is a class of approaches that deals with the problem of document classification from a theoretical point of view, making no assumption on the application of these approaches. These include statistical [9], analytical [10], information retrieval [11] and information fusion [12] approaches. The second class deals with techniques that are focused on specific applications, such as baseball program summaries [13], clinical data visualization [14] and web browsing on handheld devices [15]. [16] reports a comprehensive review.

The approach presented in this chapter produce a set of items, but involves improvements over the simple set of words process in two means. Actually, we go beyond the level of keywords providing conceptual descriptions from concepts identified and extracted from the text. We propose a practical approach for extracting the most relevant keywords from the forum threads to form a summary without assumption on the application domain and to subsequently find out concepts from the keyword extraction based on statistics and synsets extraction. Then semantic similarity analysis is conducted between keywords to produce a set of semantic relevant concepts summarizing actual forum significance.

In order to substitute keywords with univocal concepts we have to build a process called Word Sense Disambiguation (WSD). Given a sentence, a WSD process identifies the syntactical categories of words and interacts with an ontology both to retrieve the exact concept definition and to adopt some techniques for semantic similarity evaluation among words. We use MorphAdorner [17] that provides facilities for tokenizing text and WordNet [18], one of the most used ontology in the Word Sense Disambiguation task.

The methodology used in this application is knowledge-based, it uses Wikipedia as a base of information with its extensive network of cross-references, portals, categories and info-boxes providing a huge amount of explicitly defined semantics.

To extract and access useful information from Wikipedia in a scalable and timely manner we use the Wikipedia Miner toolkit [http://wikipedia-miner.sourceforge.net/] including

scripts for processing Wikipedia dumps and extracting summaries such as the link graph and category hierarchy.

4. Related works in automatic text summarization

A variety of document summarization approaches have been developed recently. The paper [19] reviews leading notions and developments, and seeks to assess the state of the art for this challenging task. The review shows that some useful summarizing for various purposes can already be done but also, not surprisingly, that there is a huge amount more to do both in terms of semantic analysis and capturing the main ideas, and in terms of improving linguistic quality of the summaries. A further overview on the latest techniques related to text summarization can be found in [20]. Generally speaking, the summarization methods can be either extractive or abstractive. Extractive summarization involves assigning relevant scores to some units (e.g. sentences, paragraphs) of the document and extracting the sentences with highest scores, while abstraction summarization involves paraphrasing sections of the source document using information fusion, sentence compression and reformulation [21]. In general, abstraction can condense a text more strongly than extraction, but the required natural language generation technologies are harder to develop representing a growing field.

Sentence extraction summarization systems take as input a collection of sentences and select some subset for output into a summary. The implied sentence ranking problem uses some kind of similarity to rank sentences for inclusion in the summary [22].

For example, MEAD (http://www.summarization.com/mead/) is based on sentence extraction. For each sentence in a cluster of related documents, MEAD computes three features and uses a linear combination of the three to determine what sentences are most salient. Sentence selection is constrained by a summary length threshold, and redundant new sentences avoided by checking cosine similarity against prior ones.

Extractive summarizers can be based on scoring sentences in the source document. For example, [23] consider each document as a sequence of sentences and the objective of extractive summarization is to label the sentences in the sequence with 1 and 0 (summary or non-summary sentence).

The summarization techniques can also be classified into two groups: supervised and unsupervised techniques. In the first case they rely on pre-existing document-summary pairs, while in the second, they are based on properties and heuristics derived from the text. Supervised extractive summarization techniques treat the summarization task as a two-class classification problem at the sentence level, where the summary sentences are positive samples while the non-summary sentences are negative samples. After representing each sentence by a vector of features, the classification function can be trained in two different manners [24]. Many unsupervised methods have been developed by exploiting different features and relationships of the sentences.

Furthermore, summarization task can also be categorized as either generic or query-based. A query-based summary presents the information that is most relevant to the given queries while a generic summary gives an overall sense of the document content [21].

Text summarization can also be classified on the basis of volume of text documents available distinguishing between single document and multi-document text summarization

techniques. The article [25] presents a multi-document, multi-lingual, theme-based summarization system based on modeling text cohesion (story flow). In this paper a Naïve Bayes classifier for document summarization is also proposed. Also in [26] we can find an analysis of multi-document summarization in scientific corpora.

Finally, automatic document summarization is a highly interdisciplinary research area related with computer science, multimedia, statistics, as well as cognitive psychology. In [27] they introduce an intelligent system based on a cognitive psychology model (the event-indexing model) and the roles and importance of sentences and their syntax in document understanding. The system involves syntactic analysis of sentences, clustering and indexing sentences with five indices from the event-indexing model, and extracting the most prominent content by lexical analysis at phrase and clause levels.

5. Applications

Here we summarize a number of real-world applications which might benefit from WSD and on which experiments have been conducted [28].

5.1 Information Retrieval (IR)

Search engines do not usually use explicit semantics to prune out documents which are not relevant to a user query. An accurate disambiguation both of the document base and of the query words, would allow it to eliminate documents containing the same words used with different meanings (thus increasing precision) and to retrieve documents expressing the same meaning with different wordings (thus increasing recall).

Most of the early work on the contribution of WSD to IR resulted in no performance improvement also because only a small percentage of query words are not used in their most frequent (or predominant) sense, indicating that WSD must be very precise on uncommon items, rather than on frequent words. [29] concluded that, in the presence of queries with a large number of words, WSD cannot benefit IR. He also indicated that improvements in IR performance would be observed only if WSD could be performed with at least 90% accuracy. Encouraging evidence of the usefulness of WSD in IR has come from [30]. Assuming a WSD accuracy greater than 90%, they showed that the use of WSD in IR improves the precision by about 4.3%. With lower WSD accuracy (62.1%) a small improvement (1.73% on average) can still be obtained.

5.2 Information Extraction (IE)

In detailed application domains it is interesting to distinguish between specific instances of concepts: for example, in the medical domain we might be interested in identifying all kinds of antidepressant drugs across a text, whereas in bioinformatics we would like to solve the ambiguities in naming genes and proteins. Tasks like named-entity recognition and acronym expansion that automatically spells out the entire phrase represented (a feature found in some content management and Web-based search systems), can all be cast as disambiguation problems, although this is still a relatively new area. Acronym expansion functions for search is considered an accessibility feature that is useful to people who have difficulties in typing.

[31] proposed the application of a link analysis method based on random walks to solve the ambiguity of named entities. [32] used a link analysis algorithm in a semi-supervised approach to weigh entity extraction patterns based on their impact on a set of instances.

Some tasks at Semeval-2007 more or less directly dealt with WSD for information extraction. Specifically, the metonymy task in which a concept is not called by its own name but by the name of something intimately associated with that concept ("Hollywood" is used for American cinema and not only for a district of Los Angeles) required systems to associate the appropriate metonymy with target named entities. Similarly, the Web People Search task required systems to disambiguate people names occurring in Web documents, that is, to determine the occurrence of specific instances of people within texts.

5.3 Machine Translation (MT)

Machine translation (MT), the automatic identification of the correct translation of a word in context, is a very difficult task. Word sense disambiguation has been historically considered as the main task to be solved in order to enable machine translation, based on the intuitive idea that the disambiguation of texts should help translation systems choose better candidates. Recently, [33] showed that word sense disambiguation can help improve machine translation. In these works, predefined sense inventories were abandoned in favor of WSD models which allow it to select the most likely translation phrase. MT tools have become an urgent need also in a multilingual environment. Although there are any available tools, unfortunately, a robust MT approach is still an open research field.

5.4 Content analysis

The analysis of the general content of a text in terms of its ideas, themes, etc., can certainly benefit from the application of sense disambiguation. For instance, the classification of blogs or forum threads has recently been gaining more and more interest within the Internet community: as blogs grow at an exponential pace, we need a simple yet effective way to classify them, determine their main topics, and identify relevant (possibly semantic) connections between blogs and even between single blog posts. [34]. A second related area of research is that of (semantic) social network analysis, which is becoming more and more active with the recent evolutions of the Web.

Although some works have been recently presented on the semantic analysis of content [35], this is an open and stimulating research area.

5.5 Lexicography

WSD and lexicography (i.e., the professional writing of dictionaries) can certainly benefit from each other: sense-annotated linguistic data reduces the considerable overhead imposed on lexicographers in sorting large-scaled corpora according to word usage for different senses. In addition, word sense disambiguation techniques can also allow language learners to access example sentences containing a certain word usage from large corpora, without excessive overhead. On the other side, a lexicographer can provide better sense inventories and sense annotated corpora which can benefit WSD.

5.6 The semantic web

The Semantic Web offers a generic infrastructure for interchange, integration and creative reuse of structured data, which can help to cross some of the boundaries that Web 2.0 is facing. Currently, Web 2.0 offers poor query possibilities apart from searching by keywords or tags. There has been a great deal of interest in the development of semantic-based systems to facilitate knowledge representation and extraction and content integration [36], [37]. Semantic-based approach to retrieving relevant material can be useful to address issues like trying to determine the type or the quality of the information suggested from a personalized environment. In this context, standard keyword search has a very limited effectiveness. For example, it cannot filter for the type of information, the level of information or the quality of information.

Potentially, one of the biggest application areas of content-based exploration might be personalized searching framework (e.g., [38],[39]). Whereas today's search engines provide largely anonymous information, new framework might highlight or recommend web pages or content related to key concepts. We can consider semantic information representation as an important step towards a wide efficient manipulation and discovery of information [40], [41], [42]. In the digital library community a flat list of attribute/value pairs is often assumed to be available. In the Semantic Web community, annotations are often assumed to be an instance of an ontology. Through the ontologies the system will express key entities and relationships describing resources in a formal machine-processable representation. An ontology-based knowledge representation could be used for content analysis and object recognition, for reasoning processes and for enabling user-friendly and intelligent multimedia content exploration and retrieval.

Therefore, the semantic Web vision can potentially benefit from most of the above-mentioned applications, as it inherently needs domain-oriented and unrestricted sense disambiguation to deal with the semantics of documents, and enable interoperability between systems, ontologies, and users.

WSD has been used in semantic Web-related research fields, like ontology learning, to build domain taxonomies. Indeed, any area of science that relies on a linguistic bridge between human and machine will use word sense disambiguation.

5.7 Web of data

Although the Semantic Web is a Web of data, it is intended primarily for humans; it would use machine processing and databases to take away some of the burdens we currently face so that we can concentrate on the more important things that we can use the Web for.

The idea behind Linked Data [43] is using the Web to allow exposing, connecting and sharing linking data through dereferenceable URIs on the Web. The goal is to extend the Web by publishing various open datasets as RDF triples and by setting RDF links between data items from several data sources. Using URIs, everything can be referred to and looked up both by people and by software agents. In this chapter we focus on DBpedia [44], that is one of the main clouds of the Linked Data graph. DBpedia extracts structured content from Wikipedia and makes this information available on the Web; it uses the RDF to represent the extracted information. It is possible to query relationships and properties associated with

Wikipedia resources (through its SPARQL endpoint), and link other data sets on the web to DBpedia data.

The whole knowledge base consists of over one billion triples. DBpedia labels and abstracts of resources are stored in more than 95 different languages. The graph is highly connected to other RDF dataset of the Linked Data cloud. Each resource in DBpedia is referred by its own URI, allowing to precisely get a resource with no ambiguity. The DBpedia knowledge base is served as Linked Data on the Web. Actually, various data providers have started to set RDF links from their data sets to DBpedia, making DBpedia one of the central interlinking-hubs of the emerging Web of Data.

Compared to other ontological hierarchies and taxonomies, DBpedia has the advantage that each term or resource is enhanced with a rich description including a textual abstract. Another advantage is that DBpedia automatically evolves as Wikipedia changes. Hence, problems such as domain coverage, content freshness, machine-understandability can be addressed more easily when considering DBpedia. Moreover, it covers different areas of the human knowledge (geographic information, people, films, music, books, …); it represents real community agreement and it is truly multilingual.

6. Using Wikipedia and WordNet in our project

For the general public, Wikipedia represents a vast source of knowledge. To a growing community of researchers and developers it also represents a huge, constantly evolving collection of manually defined concepts and semantic relations. It is a promising resource for natural language processing, knowledge management, data mining, and other research areas.

In our project we used WikipediaMiner toolkit [wikipedia-miner.sourceforge.net/], a functional toolkit for mining the vast amount of semantic knowledge encoded in Wikipedia providing access to Wikipedia's structure and content, allowing terms and concepts to be compared semantically, and detecting Wikipedia topics when they are mentioned in documents. We now describe some of the more important classes we used to model Wikipedia's structure and content.

Pages: All of Wikipedia's content is presented on pages of one type or another. The toolkit models every page as a unique id, a title, and some content expressed as MediaWiki markup.

Articles provide the bulk of Wikipedia's informative content. Each article describes a single concept or topic, and their titles are succinct, well-formed phrases that can be used as non-descriptors in ontologies and thesauri. For example, the article about domesticated canines is entitled Dog, and the one about companion animals in general is called Pet. Once a particular article is identified, related concepts can be gathered by mining the articles it links to, or the ones that link to it.

The anchor texts of the links made to an article provide a source of synonyms and other variations in surface form. The article about dogs, for example, has links from anchors like canis familiaris, man's best friend, and doggy.

The subset of keywords related to each article helps to discriminate between concepts. In such a way, two texts characterized using different keywords may result similar considering underling concept and not the exact terms. We use the WordNet to perform the following

feature extraction pre-process. Firstly, we label occurrences of each word as a part of speech (POS) in grammar. This POS tagger discriminates the POS in grammar of each word in a sentence. After labeling all the words, we select those ones labeled as noun and verbs as our candidates. We then use the stemmer to reduce variants of the same root word to a common concept and filter the stop words.

WordNet is an online lexical reference system, in which English nouns, verbs, adjectives and adverbs are organized into synonym sets. Each synset represents one sense, that is one underlying lexical concept. Different relations link the synonym sets, such as IS-A for verbs and nouns, IS-PART-OF for nouns, etc. Verbs and nouns senses are organized in hierarchies forming a "forest" of trees. For each keyword in WordNet, we can have a set of senses and, in the case of nouns and verbs, a generalization path from each sense to the root sense of the hierarchy. WordNet could be used as a useful resource with respect to the semantic tagging process and has so far been used in various applications including Information Retrieval, Word Sense Disambiguation, Text and Document Classification and many others.

Noun synsets are related to each other through hypernymy (generalization), hyponymy (speciali-zation), holonymy (whole of) and meronymy (part of) relations. Of these, (hypernymy, hyponymy) and (meronymy, holonymy) are complementary pairs. The verb and adjective synsets are very sparsely connected with each other. No relation is available between noun and verb synsets. However, 4500 adjective synsets are related to noun synsets with pertainyms (pertaining to) and attra (attributed with) relations.

Articles often contain links to equivalent articles in other language versions of Wikipedia. The toolkit allows the titles of these pages to be mined as a source of translations; the article about dogs links to (among many others) chien in the French Wikipedia, haushund in German, and 犬 in Chinese.

Redirects are pages whose sole purpose is to connect an article to alternative titles. Like incoming anchor texts, these correspond to synonyms and other variations in surface form. The article entitled dog, for example, is referred to by redirects dogs, canis lupus familiaris, and domestic dog. Redirects may also represent more specific topics that do not warrant separate articles, such as male dog and dog groups.

Categories: Almost all of Wikipedia's articles are organized within one or more categories, which can be mined for hyponyms, holonyms and other broader (more general) topics. Dog, for example, belongs to the categories domesticated animals, cosmopolitan species, and scavengers. If a topic is broad enough to warrant several articles, the central article may be paired with a category of the same name: the article dog is paired with the category dogs. This equivalent category can be mined for more parent categories (canines) and subcategories (dog breeds, dog sports). Child articles and other descendants (puppy, fear of dogs) can also be mined for hypernyms, meronyms, and other more specific topics.

All of Wikipedia's categories descend from a single root called Fundamental. The toolkit uses the distance between a particular article or category and this root to provide a measure of its generality or specificity. According to this measure Dog has a greater distance than carnivores, which has the same distance as omnivores and a greater distance than animals.

Disambiguations: When multiple articles could be given the same name, a specific type of article—a disambiguation—is used to separate them. For example, there is a page entitled dog

(disambiguation), which lists not only the article on domestic dogs, but also several other animals (such as prairie dogs and dogfish), several performers (including Snoop Doggy Dogg), and the Chinese sign of the zodiac. Each of these sense pages have an additional scope note; a short phrase that explains why it is different from other potential senses.

Anchors, the text used within links to Wikipedia articles, are surprisingly useful. As described earlier, they encode synonymy and other variations in surface form, because people alter them to suit the surrounding prose. A scientific article may refer to canis familiaris, and a more informal one to doggy. Anchors also encode polysemy: the term dog is used to link to different articles when discussing pets, star signs or the iconic American fast food. Disambiguation pages do the same, but link anchors have the advantage of being marked up directly, and therefore do not require processing of unstructured text. They also give a sense of how likely each sense is: 76% of Dog links are made to the pet, 7% to the Chinese star sign, and less than 1% to hot dogs.

Wikipedia itself is, of course, one of the more important objects to model. It provides the central point of access to most of the functionality of the toolkit. Among other things, here you can gather statistics about the encyclopedia, or access the pages within it through iteration, browsing, and searching.

We used RitaWn [http://www.rednoise.org/rita/wordnet] to query WordNet.

7. System architecture

This section describes the logical structure of the project, describing software components (Figure1) and database (Figure 2) that allow the system to carry out its task.

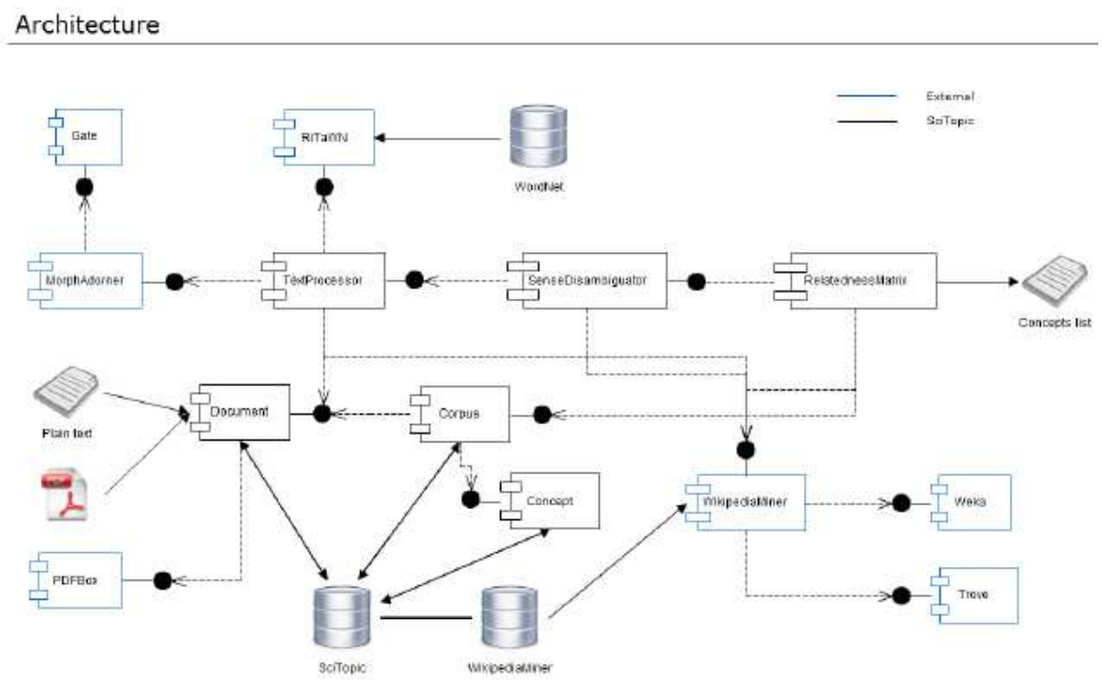


Fig. 1. system architecture

The database maintains documents in plain text and organize them in one or more possibly *Corpus*, a set of documents linked together in a logical manner, for example, by dealing with the topic.

We associate a frequency list of concepts for each corpus (and therefore a set of documents), that is how many times a particular concept is repeated within all the documents of the corpus. A concept corresponds exactly to a Wikipedia article, thus creating a relationship between our project and WikipediaMiner database, more precisely, between the ConceptFrequency and Page tables. These statistics, as we shall see below, are used to better identify those concepts that define a document between similar.

To abstract and manage documents and databases we have created two components, *Corpus* and *Document*, which act as an interface between the other components of the application and the database. They provide editing, creation and deletion functions to facilitate the extraction of content representing the input to all other phases of the main process.

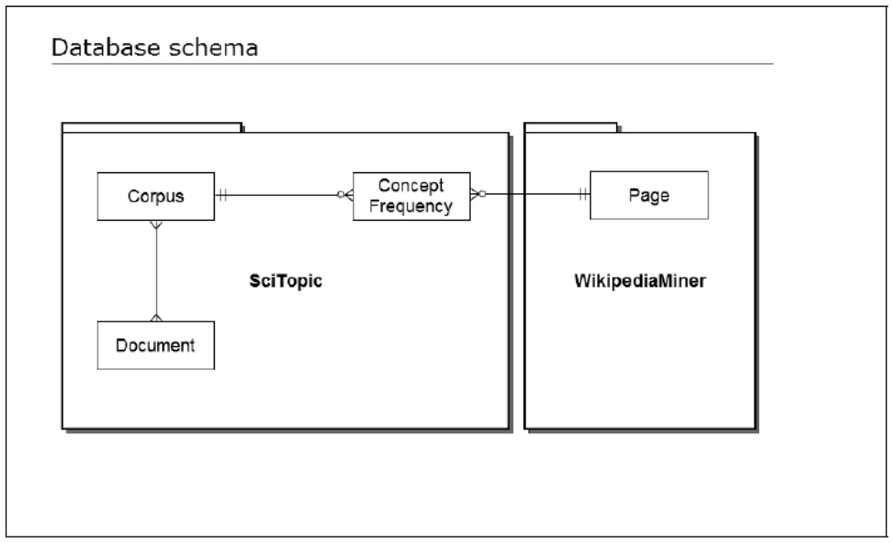


Fig. 2. system database

Starting from a document and a corpus to which it is associated, we proceed to a series of transformations on the text in order to filter all unnecessary components leaving only a set of names in basic form useful for later phases. The component that performs this task is the *TextProcessor* and is configurable, allowing the user to define the desired level of filtering.

The output of the previous phase is used for the disambiguation task, carried out by the component *SenseDisabiguator*; the system maps the most appropriate Wikipedia article to each term or, if this is not possible, it eliminates the term considered unknown. The result is a series of concepts, that is Wikipedia articles.

The component RelatednessMatrix uses this list of concepts to establish the importance of each of them within the context. In particular, the system performs the sum of the degree of relationship between a concept and all the others and evaluates this amount depending on the TFxIDF. So doing, the system associates a weight to each concept, reaching the objective of obtaining the *n* that best define the content of the input document.

7.1 Text processor

Starting from a document and a corpus to which it is associated, *TextProcessor* performs transformations on the text in order to filter all unnecessary components leaving only a set of names in basic form useful for later phases (see Figure 3).

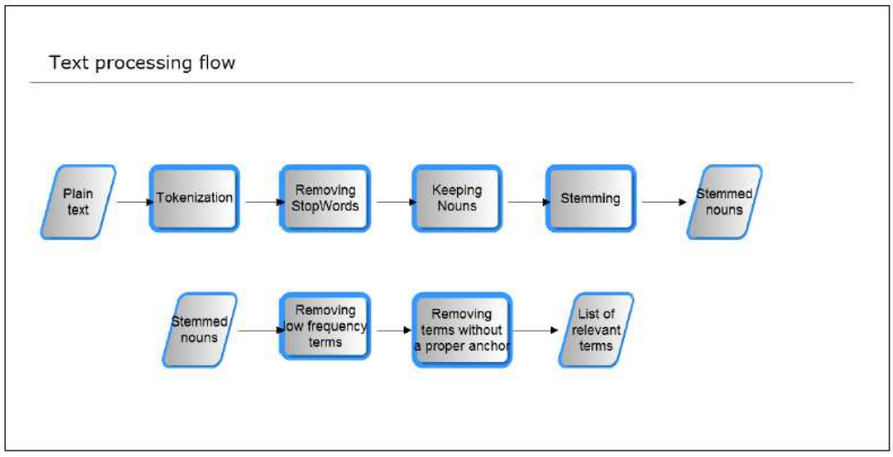


Fig. 3. The text processing flow

The disambiguation process is expensive and it is therefore appropriate to delete any irrelevant term; with this intention the *TextProcessor* module removes both all the remaining words with a frequency less than a given threshold and all those that do not correspond to an appropriate Wikipedia anchor, that is an anchor that has no meaning with probability greater than a defined minimum.

Summarizing, *TextProcessor* has two parameters that affect the selectivity of performed functions: the minimum frequency and the minimum probability of the Wikipedia senses (articles).

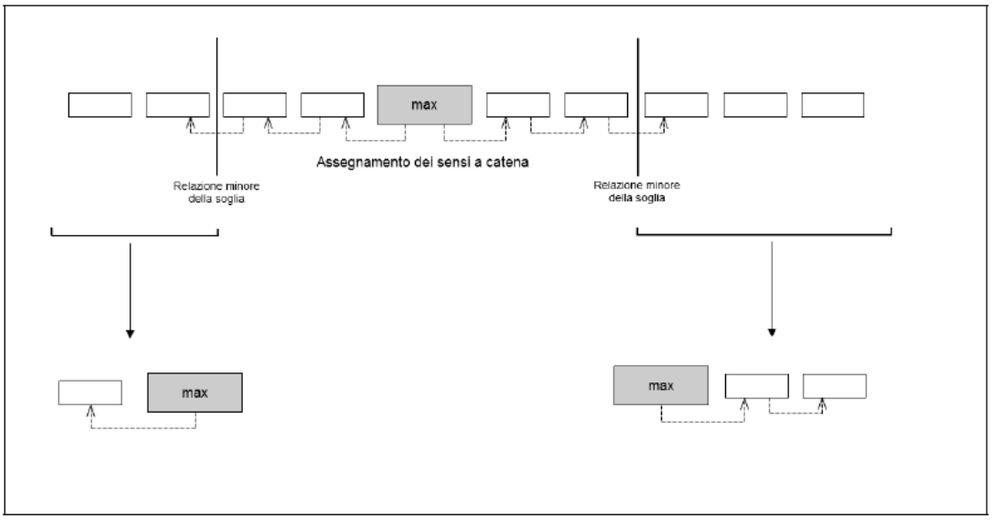


Fig. 4. The SenseDisambiguator recursive procedure

7.2 Sense disambiguator

It is the component that assigns to a specific list of terms (see Figure 4) the most appropriate sense. It is achieved by a recursive procedure (see Figure 5) that takes a list of terms and recursively splits it into slices; for each part it defines a main sense from which disambiguate the others.

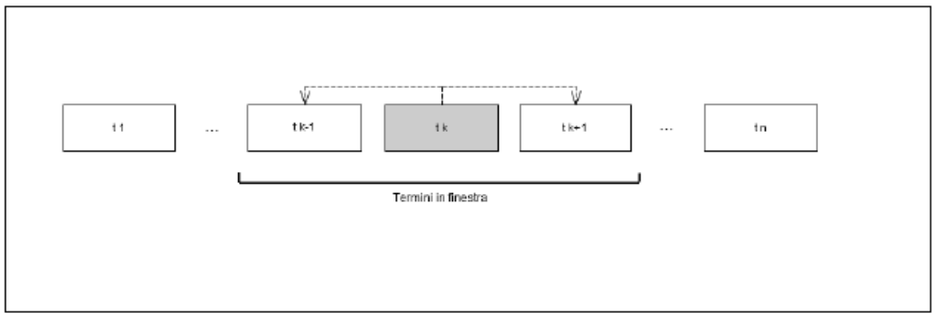


Fig. 5. The sense score is calculated based on the term senses belonging to a window

The pseudo-code of the disambiguation procedure is showed in the following listing.

Listing 6.1: Pseudocodice della procedura di disambiguazione

```
1
2 WSD(termini)
3
4 //Calcolo dei punteggi dei sensi dei termini
5 foreach(Termine t1 in termini){
6     foreach(Senso s1 in sensi(t1)){
7         foreach(Termine t2 in finstra(t1)){
8             foreach(Senso s2 in sensi(t2)){
9                 Aggiorna punteggio(s1,t1)
10            }
11        }
12    }
13 }
14
15 AssegnaSensi(termini, 1, n)
```

7.3 Relatedness matrix

The RelatednessMatrix has the task of both building a relationship matrix using all elements of a given list of senses and of providing various ways of extracting information. It is the component used in the final phase, that is, given a list of senses it extracts the most relevant.

Listing 6.2: Pseudocodice della procedura di assegnamento dei sensi

```
1
2 AssegnaSensi(termini, inizio, fine)
3
4 //Termine con senso a punteggio massimo
5 smax := {s:max{punteggio(s,t)}}
6 tmax := {t:smax in t}
7
8 //Assegno i sensi dei termini a sinistra di tmax
9 foreach(Termine t1 in termini a sinistra){
10     if(Esiste almeno un senso che ha una relazione sufficiente con
           il senso assegnato al termine precedente){
11         Assegna al termine t1 il senso s che massimizza relazione(s
               , s_precedente)
12     }else{
13         // Ricorsione
14         AssegnaSensi(termini, inizio, t1)
15         break
16     }
17 }
18
19 Unisci liste di sensi
20
21 //Assegno i sensi dei termini a destra di tmax
22 foreach(Termine t1 in termini a destra){
23     if(Esiste almeno un senso che ha una relazione sufficiente con
           il senso assegnato al termine precedente){
24         Assegna al termine t1 il senso s che massimizza punteggio(s
               , s_precedente)
25     }else{
26         // Ricorsione
27         AssegnaSensi(termini, t1, fine)
28         break
29     }
30 }
31
32 Unisci liste di sensi
33
34 return
```

8. Considerations

The work described in this chapter represents some initial steps in exploring automatic concept extraction in semantic summarization process. It could be considered as one possible instance of a more general concept concerning the transition from the Document Web to the Document/Data Web and the consequent managing of these immense volumes of data.

Summarization can be evaluated using intrinsic or extrinsic measures; while the first one methods attempt to measure summary quality using human evaluation, extrinsic methods measure the same through a task-based performance measure such the information retrieval-oriented task. In our experiments we utilized intrinsic approach analyzing [45] as document and [45] [46] [47] [48] [49] [50] [51] [52] [53] [54] [55] as corpus.

This experiment is to evaluate the usefulness of concept extraction in summarization process, by manually reading whole document content and comparing with automatic extracted concepts. The results show that automatic concept-based summarization produces useful support to information extraction. The extracted concepts represent a good summarization of document contents.

For example, we evaluated the influence of chosen window size using 605 terms to be disambiguated. The results are showed in Table 1.

Window	Copertura (C)	Precisione (P)	C%	P%
4	438	268	72.39	61.18
8	461	357	76.19	77.44
12	470	355	77.68	75.53
16	475	369	78.51	77.68
20	480	362	79.33	75.41

Table 1. Change in precision and recall as a function of window size.

Using the best choice of parameter values we obtain the following percentages in precision and recall.

window	minScore	minRelatednessToSplit
8	0.18	0.2

Copertura (C)	Precisione (P)	C%	P%
444	358	73.38	80.63

Table 2. Change in precision and recall using the showed set of parameter values.

Finally, given the document [MW08], Table 3 shows the ten most representative articles automatically extracted from the system.

While the initial results are encouraging, much remains to be explored. For example, many disambiguation strategies with specific advantages are available, so designers now have the possibility of deciding which new features to include in order to support them, but it is particularly difficult to distinguish the benefits of each advance that have often been shown independent of others.

Listing 6.16: I primi dieci articoli che rappresentano il documento [MW08].

```
1 Language
2 System
3 Category theory
4 Knowledge
5 Word
6 Datum (geodesy)
7 Concept
8 WordNet
9 Application software
10 Accuracy and precision
```

Table 3. Automatically extracted articles representing [MW08]

It would also be interesting to apply the showed method using a different knowledge base, for example YAGO (but always derived from Wikipedia) and use a different measure of relationship between concepts considering not only the links belonging to articles but also the entire link network. That is, considering Wikipedia as a graph of interconnected concepts, we could exploit more than one or two links.

9. References

[1] Henze N., Dolog P., Nejdl W.: Reasoning and Ontologies for Personalized E-Learning in the Semantic Web, Educational Technology & Society, 7 (4), 82-97 (2004)

[2] Bighini C., Carbonaro A.: InLinx: Intelligent Agents for Personalized Classification, Sharing and Recommendation, International Journal of Computational Intelligence. International Computational Intelligence Society. 2 (1), (2004)

[3] Pickens, J., Golovchinsky, G., Shah, C., Qvarfordt, P., and Back, M. Algorithmic Mediation for Collaborative Exploratory Search. To appear in Proceedings of SIGIR

[4] Freyne J., Smyth B.: Collaborative Search: Deployment Experiences, in The 24th SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence. Cambridge, UK, pp. 121-134 (2004)

[5] Calic J., Campbell N., Dasiopoulou S. and Kompatsiaris Y.: A Survey on Multimodal Video Representation for Semantic Retrieval, in the Third International Conference on Computer as a tool, IEEE (2005)

[6] Carbonaro A., Defining Personalized Learning Views of Relevant Learning Objects in a Collaborative Bookmark Management System, In Z. Ma (Ed.), Web-based Intelligent ELearning Systems: Technologies and Applications (pp. 139-155). Hershey, PA: Information Science Publishing, (2006)

[7] Bloehdorn S., Petridis K., Simou N., Tzouvaras V., AvrithisY., Handschuh S., Kompatsiaris Y., Staab S., Strintzis M. G.: Knowledge Representation for Semantic Multimedia Content Analysis and Reasoning, in Proceedings of the European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology, (2004)

[8] White R. W. and Roth R., Exploratory Search: Beyond the Query-Response Paradigm. Morgan & Claypool, (2009)

[9] McKeown, K., Barzilay, R., Evans, D., Hatzivassiloglou, V., Kan, M., Schiffman, B., and Teufel, S. "Columbia Multi- Document Summarization: Approach and Evaluation". Workshop on Text Summarization, 2001.

[10] Brunn, M., Chali, Y., and Pinchak. C. "Text Summarization Using Lexical Chains". Work. on Text Summarization. 2001.

[11] Aho, A., Chang, S., McKeown, K., Radev, D., Smith, J., and Zaman, K. "Columbia Digital News Project: An Environment for Briefing and Search over Multimedia". Information J. Int. J. on Digital Libraries, 1(4):377-385. 1997.

[12] Barzilay, R., McKeown, K. and Elhadad, M. "Information fusion in the context of multi-document summarization". In Proc. of ACL'99, 1999.

[13] Yong Rui, Y., Gupta, A., and Acero, A. "Automatically extracting highlights for TV Baseball programs". ACM Multimedia, Pages 105-115, 2000.

[14] Shahar, Y. and Cheng, C. "Knowledge-based Visualization of Time Oriented Clinical Data". Proc AMIA Annual Fall Symp., pages 155-9, 1998.

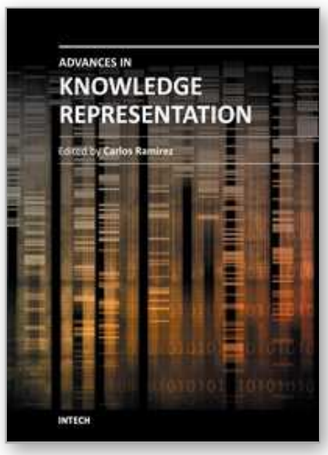
[15] Rahman, A, H. Alam, R. Hartono and K. Ariyoshi. "Automatic Summarization of Web Content to Smaller Display Devices", 6th Int. Conf. on Document Analysis and Recognition, ICDAR01, pages 1064-1068, 2001.

[16] NIST web site on summarization: http://wwwnlpirnist.gov/projects/duc/pubs.html, Columbia University Summarization Resources (http://www.cs.columbia.edu/~hjing/ summarization.html) and Okumura-Lab Resources (http://capella.kuee.kyoto-u.ac.jp/index_e.html).

[17] Burns, Philip R. 2006. MorphAdorner: Morphological Adorner for English Text. http://morphadorner.northwestern.edu/morphadorner/textsegmenter/.

[18] Fellbaum, C., ed (1998), WordNet: An Electronic Lexical Database. MIT Press, Cambridge, Mass
[19] Jones, K. S. (2007). Automatic summarizing: The state of the art. Information Processing and Management, 43, 1449 – 1481
[20] Proceedings of the Workshop on Automatic Text Summarization 2011 Collocated with Canadian Conference on Artificial Intelligence St. John's, Newfoundland and Labrador, Canada May 24, 2011
[21] Wan, X. (2008). Using only cross-document relationships for both generic and topic-focused multi-document summarizations. Information Retrieval, 11, 25-49
[22] Alguliev, R. M., & Alyguliev, R. M. (2007). Summarization of text-based documents with a determination of latent topical sections and information-rich sentences. Automatic Control and Computer Sciences, 41, 132 – 140
[23] Shen, D., Sun, J. -T., Li, H., Yang, Q., & Chen, Z. (2007). Document summarization using conditional random fields. In Proceedings of the 20th international joint conference on artificial intelligence (IJCAI 2007), January 6 – 12 (pp. 2862 – 2867) Hyderabad, India.
[24] Mihalcea, R., & Ceylan, H. (2007). Explorations in automatic book summarization. In Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL 2007), 28 – 30 June (pp. 380 – 389) Prague, Czech Republic.
[25] Fung, P., & Ngai, G. (2006). One story, one flow: Hidden Markov story models for multilingual multidocument summarization. ACM Transaction on Speech and Language Processing, 3, 1 – 16.
[26] Ozge Yeloglu, Evangelos Milios, Nur Zincir-Heywood, SAC'11 March 21-25, 2011, TaiChung, Taiwan. Copyright 2011 ACM 978-1-4503-0113-8/11/03
[27] Guo, Y., & Stylios, G. (2005). An intelligent summarization system based on cognitive psychology. Information Sciences, 174, 1 – 36.
[28] Navigli, R. 2009.Word sense disambiguation: A survey. ACM Comput. Surv. 41, 2, Article 10 (2009)
[29] Sanderson, M. 1994. Word sense disambiguation and information retrieval. In Proceedings of the Special Interest Group on Information Retrieval (SIGIR, Dublin, Ireland). 142–151.
[30] Stokoe, C., Oakes, M. J., and Tait, J. I. 2003. Word sense disambiguation in information retrieval revisited. In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (Toronto, Onto., Canada). 159–166
[31] Malin, B., Airoldi, E., and Carley, K. M. 2005. A network analysis model for disambiguation of names in lists. Computat. Math. Organizat. Theo. 11, 2, 119–139.
[32] Hassan, H., Hassan, A., and Noeman, S. 2006. Graph based semi-supervised approach for information extraction. In Proceedings of the TextGraphs Workshop in the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT/NAACL, New York, NY). 9–16
[33] Chan, Y. S., Ng, H. T., and Chiang, D. 2007a. Word sense disambiguation improves statistical machine translation. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (Prague, Czech Republic). 33–40.
[34] A. Carbonaro "Forum Summarization to Support Tutor and Teacher in group interaction management" Lytras M.D., De Pablos P.O., Damiani E. (eds.) Semantic Web Personalization and Context Awareness: Management of Personal Identities

and Social Networking, Information Science Reference, chapter 3, pp. 22-31, IGI-Global. 2011, ISBN 978-1-61520-921-7
[35] Semantic Analysis: A Practical Introduction (Oxford Textbooks in Linguistics), Cliff Goddard, 2011
[36] Henze N., Dolog P., Nejdl W.: Reasoning and Ontologies for Personalized E-Learning in the Semantic Web, Educational Technology & Society, 7 (4), 82-97 (2004)
[37] Bighini C., Carbonaro A.: InLinx: Intelligent Agents for Personalized Classification, Sharing and Recommendation, International Journal of Computational Intelligence. International Computational Intelligence Society. 2 (1), (2004)
[38] Pickens, J., Golovchinsky, G., Shah, C., Qvarfordt, P., and Back, M. Algorithmic Mediation for Collaborative Exploratory Search. To appear in Proceedings of SIGIR
[39] Freyne J., Smyth B.: Collaborative Search: Deployment Experiences, in The 24th SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence. Cambridge, UK, pp. 121-134 (2004)
[40] Calic J., Campbell N., Dasiopoulou S. and Kompatsiaris Y.: A Survey on Multimodal Video Representation for Semantic Retrieval, in the Third International Conference on Computer as a tool, IEEE (2005)
[41] Carbonaro A., Defining Personalized Learning Views of Relevant Learning Objects in a Collaborative Bookmark Management System, In Z. Ma (Ed.), Web-based Intelligent ELearning Systems: Technologies and Applications (pp. 139-155). Hershey, PA: Information Science Publishing, (2006)
[42] Bloehdorn S., Petridis K., Simou N., Tzouvaras V., AvrithisY., Handschuh S., Kompatsiaris Y., Staab S., Strintzis M. G.: Knowledge Representation for Semantic Multimedia Content Analysis and Reasoning, in Proceedings of the European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology, (2004)
[43] Bizer, C. Heath T., and T. Berners-Lee. Linked data - the story so far. International Journal on Semantic Web and Information Systems, 5(3):1, (2009)
[44] Bizer C., Lehmann J., Kobilarov G., Auer S., Becker C., Cyganiak R., and S. Hellmann. Dbpedia – a crystallization point for the web of data. Web Semantics: Science, Services and Agents on the World Wide Web, (2009)
[45] D. Milne and I. H. Witten. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. 2008
[46] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: A large ontology from wikipedia and wordnet. 2008
[47] J. Yu, J. A. Thom, and A. Tam. Ontology evaluation using wikipedia categories for browsing. 2007
[48] A. Kittur, E. H. Chi, and B. Suh. What's in wikipedia? Mapping topics and conflict using socially annotated category structure. 2009.
[49] T. Zesch, C. M¨uller, and I. Gurevych. Extracting lexical semantic knowledge from wikipedia and wiktionary. 2008.
[50] E. Wolf and I. Gurevych. Aligning sense inventories in Wikipedia and wordnet. 2010.
[51] D. Milne and I. H. Witten. An open-source toolkit for mining wikipedia. 2009
[52] P. Sch¨onhofen. Identifying document topics using the Wikipedia category network. 2009
[53] O. Medelyan, I. H. Witten, and D. Milne. Topic indexing with wikipedia. 2008
[54] H. Amiri, M. Rahgozar, A. A. Ahmad, and F. Oroumchian. Query expansion using wikipedia concept graph. 2008
[55] R. Mihalcea and A. Csomai. Wikify! linking documents to encyclopedic knowledge. 2007



Advances in Knowledge Representation

Edited by Dr. Carlos Ramirez

ISBN 978-953-51-0597-8
Hard cover, 272 pages
Publisher InTech
Published online 09, May, 2012
Published in print edition May, 2012

Advances in Knowledge Representation offers a compilation of state of the art research works on topics such as concept theory, positive relational algebra and k-relations, structured, visual and ontological models of knowledge representation, as well as detailed descriptions of applications to various domains, such as semantic representation and extraction, intelligent information retrieval, program proof checking, complex planning, and data preparation for knowledge modelling, and a extensive bibliography. It is a valuable contribution to the advancement of the field. The expected readers are advanced students and researchers on the knowledge representation field and related areas; it may also help to computer oriented practitioners of diverse fields looking for ideas on how to develop a knowledge-based application.

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen