Journal of Universal Language 2
March 2001, 34-50

Artificial Intelligence and Artificial Language: Agents in Information Architecture for Intelligent Distributed Multilingual Document Retrieval Service

Key-sun Choi
Korea Advanced Institution of Science and Technology

Abstract

"Intelligence**"** is made from context-adaptable search mechanisms over domain-specific contexts in viewpoint of artificial intelligence. In this paper, "communication" between two parties is accomplished by this "intelligence". "Understanding" states of two parties are agreed to be a basis of communication. However, the communicated contents vary with their level of understanding for the current contexts in each communication party. Each level of understanding is based on different levels of contents. Each level of contents is represented by different "language". This paper tries to explain the information architecture of contents communication. If there is a gap of knowledge level between two communication parties, their level of communication must be moved to the more deep understanding level. In other words, speakers (or authors) of contents must give more information to the hearers (or readers). If speakers cannot give more information, they may be supplied from other sources (or agents) that arc distributed in an accessible network. An example of this architecture is realized as an intelligent distributed multilingual document retrieval service.

1. Introduction

The document retrieval service assumes that the right information

is passed to the right location via the right paths. The right information for a given location is gathered from the collection of documents by using some methods of search, extraction, filtering, browsing, and so on. Here, a “location” means a user, an agent or a computer who requested information by a certain query statement. A unit of “information” comes from “documents”. A document is a surface of information which represents that document. A unit of information is gathered from a set of documents under a certain constraint of view. A “view” of location (or author) makes different the expression of information. Because every author has different view of expression, each document as a writing result of author is embodied in a different way even if given the same contents of information. That is why the same information can be extracted from a collection of documents each of which has different surface expression.

The right path of information is to find the most efficient and shortest way from the location to the information (or document) or vice versa. If we stand in a standpoint of “information,” information trips from one document to another, and finally arrives at a right location. We call it “information extraction (or retrieval)” or “document retrieval” according to the viewpoint of information and document respectively. On the other hand, standing from a “location,” we navigate from one document to another by browsing them until a right information is found in a document. That is called “information searching” or "navigation" of information space (or document space).

“Information flow” is a term to describe both information extraction and information navigation. “Intelligent" document retrieval is to get the efficient way of information flow based on some “understanding” of documents. An intelligent way of document retrieval does not come from a non-understanding of those documents. If documents arc understandable to a location-whether the location is a person, an agent or a computer, we can find the right path to the right documents for the given query of right location.

“Understanding" states reside in both documents and location. If a document is said to be understandable, that is written in well-formed sentences, well-presented styles, well-structured information, well-expressed contents of scenario based on the knowledge of the assumed readers, which is well-fitted to the assumed readers' viewpoints, well-anchored to the assumed readers' situation, and reasoning capability. If a location is not enough to understand the document, the location can ask again the author s location for more explanation.

If some readers understand a document but others cannot, those locations are located in the different knowledge level. (Here, we did not define what is “knowledge level”.) In other words, there is a knowledge gap between those locations.

Now, we would like to return to our problem “efficient document retrieval.” The following is a summary of what we discussed up to now. Efficient document retrieval assumed “understanding” states, which assumed a so-called “complete document state. A complete document can be accomplished by integrating both document completeness and location’s understandability. Our problem turns out to be what a complete document consists of. A model document architecture will be presented to explain the "understanding” of given documents in the next section.

Document architecture is to represent a complete understanding of the state of the document. As stated in the above, á complete document is a result of compensatory integration of document and location. The restoring process toward complete document needs several processes and resources. This restoring process is accomplished not limited to one location. Multiple locations participate in the restoring process by manner of cooperation or competition. We call it information architecture.

“Multi-lingual” service is explained by “understandability” in document architecture. Furthermore, multi-lingual processing (or translation) is just the restoring process on the information architecture.

Finally, the implementation of document architecture assumes the

standardized specification of each of its layers. Here, the "information interchange format" means the specification of communication between each layer of document architecture and information architecture.

2. Document Architecture for Intelligent Document Retrieval

2.1. Document Architecture and Understanding of Document

Document architecture contains everything for understanding a document. For example, document architecture can consist of five layers inside a document: character (including any object), presentation layout, presented data information structure and knowledge. The first two layers (character and layout) are categorized into the "surface" representation of document. The layer of presented data and information structure can be seen as a syntactic structure, and as a part of the information structure and the knowledge as semantics of a document.

A document is written in a character (or figure, etc.) and in a layout structure (for example, title, paragraph, etc.) When the author writes a document, s/he knows about written characters and its layout structure. If he types it in a computer, his typing is stored in a standardized character code. These are surface representation of documents, for every document is written in a character and in a layout. The reader is assumed to recognize them. Otherwise, a process is invoked to make the reader understand them. For example, "multi-lingual" processing is invoked which will be explained later.

However, the more understanding is involved, the more structure of data should be recognized. For example, the link between a figure and its related text units is such a syntactic structure between data in the document. Whenever a reader wants to understand a document, the first job is to find such a syntactic linkage inside of document, for example, links among text units, footnotes, references, figures

and so on.

Moreover, a reader (location) wants to read a document, it should understand the document's linguistic layers: their morphological, syntactic and semantic structure. Without such understanding, a reader cannot understand sentences in document. Finally, a complete document requires knowledge of terminology and then the domain knowledge where the document is written on.

To understand a document is that the reader's location has the capability to recognize every layer of document architecture of that document. Reader's capabilities are assumed to recognize the character, document layout, structure of data, structure of information (linguistic information), terminology and domain knowledge. Consider that document retrieval is based on the understanding of documents. In other words, document retrieval is a kind of communication between authors of documents and readers who want to search and read his necessary documents among the collection of documents. Document retrieval is a kind of "communication" between information producer (writer) and consumer (reader). The action of retrieval involves the process of communication. Communication assumes the same level of knowledge between both sides, for communication is invoked with understanding. Communication with false understanding follows from the different knowledge level. We call that "communication bottleneck" which entails knowledge gaps. The problem is how to overcome such knowledge gaps by reducing the communication bottlenecks. Consider again the knowledge state of author and reader locations of document.

The author location of document understands fully its written document. It knows every knowledge level of document architecture inside of that document. The author also assumes some right knowledge level of reader when it writes the document. Whenever such expectation is not fulfilled, the communication bottleneck state holds.

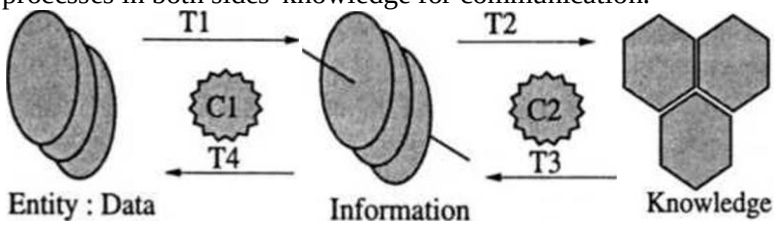
Next, let's stand on the side of reader. For a given document, a reader gets to understand it by using reader s own local knowledge.

If the reader is a human, the human's memory will be used; otherwise, if the reader is a computer, its program processes is based on its local database. Unless the reader understands fully, they use other location's knowledge: for example, dictionary or encyclopedia. If the reader is a computer, such public knowledge is written in a machine-understandable form instead of human-readable form on paper. The next step when they are not understandable yet, it that the reader asks questions to the author location. The author then responds by an appropriate answer. Such answer may fill the gap in document architecture on the reader's side. Feedback information is helpful to understand the author's intention.

2.2. Complete Document and Document Interchange Format

"Complete document" requires the description of knowledge levels inside of documents under the concept of document architecture. Whenever communication invokes, both sides of documents assumes that they recover the complete document from the surface structure of document, that is, author's writing. The recovery of complete document is the assumption of right communication. We call the specification of document architecture as "document interchange format".

A complete document is recovered from both sides' knowledge. Because both sides cooperate for reader location's understanding, they are compensating each other. Furthermore, they are also competing for the completeness. The next section will present a concept "information architecture" to clarify the entities and processes in both sides' knowledge for communication.



♦Transformation:
-T1 : Information Extraction
-T1: Knowledge Extraction
-T1: Idea Generation
-T1: Data Generation

♦Constraint:
- Cl: View, C2 : Situation

Figure 1. Configuration of Information Architecture

2.3. Information Architecture

2.3.1. Definition of Information Architecture

"Information architecture" is introduced to refine the processes under the document architecture. The configuration of information architecture (Figure I) consists of entities, transformations and constraints between transformations. Entities are data, information, and knowledge. The transformation between entities involves four processes: information extraction, knowledge acquisition, idea generation and data presentation. The constraints are two-kinds: view and situation.

The first entity of information architecture is "data". Data is just the surface of document architecture. The examples are text, word-processing output or multimedia document. The author writes data, not information in the form of document. The second entity is "information" which is the result of structuring of data. When a document is converted to its state of information, surface units in the document have cross-referential links between them. Linguistic tags annotate linguistic units. Examples of information are hypertext in a sense of structured data, and morphological/syntactic lagged sequence of units of sentences as a processed data. That is moved toward understanding. The third entity is "knowledge". It has its normalized form. Such knowledge includes terminological knowledge, domain knowledge and so on.

Information is extracted from data by cutting off the view of data presentation, information extraction" is a transformation from data

to information. Contents of many documents are summarized into one unit of information. Summarization is also one of processes of information extraction. “View”, is analogous to a “clothes” of information to be displayed to the reader of document as data. Views of authors force the same information to be written in a different presentation of documents. When a reader is a child, the document is written in a very simple manner of writing and that document contains lots of figures. Document for an expert consists of many formal formulas to convey very technical facts. Data presentation depends on the “view” of “location”. The process of “data generation” or “data generation" or “data presentation" from information to data is to produce good-looking documents depending on level of readers.

Knowledge is formalized from information after normalizing by “situation” factor. That process is called "knowledge extraction". Traditionally, terms like "learning” or “knowledge acquisition" have been used. However, those latter terms were used for the direct transformation from data to knowledge. The reverse process of knowledge extraction, “idea generation, creates information appropriate to a given situation. Situation anchors a time parameter onto a unit of knowledge and generates instances of knowledge as information. A unit of information is generated by a given situation. Here, a collection of knowledge is integrated to generate one unit of information. This process “idea generation” helps a user to make an idea. However, because such an idea is not in a final presented form of data, the idea representation may not be understandable to other location at the standpoint of “data".

2.3.2. Normalized Entity of Information Architecture

Each entity of information architecture can claim to have its optimal specification. Each entity is embodied in many applications. An application of data is "document". The optimal specifications of document include style sheet or writing scenario that is the most persuasive for a given user. They are based on the optimal

presentation method depending on a given view of location. Let us define the "well-presented data" as the optimal data form for a given user (or location). If we can find such well-presented data for generating the optimal persuasive documents, we have a goal to pursue in the study of data and information under the paradigm of information architecture.

The normalized form of information is called "well-structured information". For example, a well-linked hypertext without hassle is in normalized state. Every information unit in such hypertext can be searched in an optimal way. An entity "knowledge" has its normalized form as "well-formed representation". Normalized logical form is one of examples.

Every layer of document architecture is projected to one of entities in the information architecture. The next section returns to the discussion of document retrieval in distributed environment.

3. Information Architecture Network for Intelligent Distributed Document Retrieval

As seen in the former sections, the author has complete knowledge to understand the author s document but the author does not provide the complete document. However complete document is a prerequisite for intelligent document retrieval. The ultimate purpose of document retrieval is to achieve efficient communication between authors and readers. The problem is where the complete document can be recovered. It is claimed that the information architecture is one of solution.

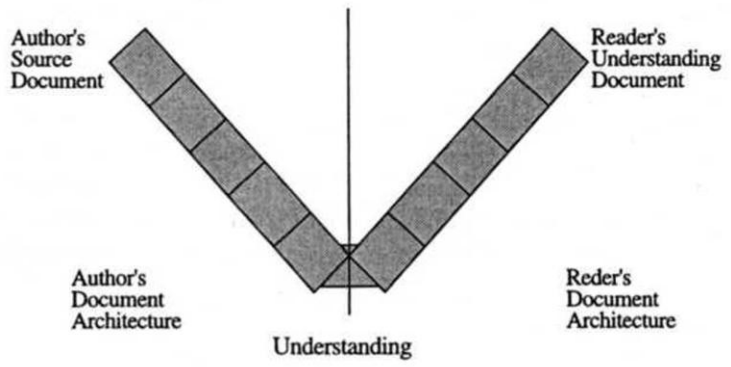


Figure 2. Distributed Information Architecture

3.1. Information Architecture Network to Recover the Document Architecture

The authors have complete knowledge to understand their writings, but they do not provide the complete documents. The source to fill up the blanks toward complete document is not in the location of author, but in the analyzing components of information architecture. That is, such knowledge location is different from the author's location. Every knowledge in the information architecture network has various physical forms. If the location is in a computer, such physical form is in machine-readable form, and they are located in either a network or its own storage. Information architecture of author's side constructs and recovers the complete document as in the left side of Figure 2. On the other hand, the reader's understanding is based on the complete document in the reader's side as in the right side of Figure 2.

The dynamics of information architecture network is different from the static view in Figure 2. The document architecture of reader and writer are compensating and competing objects. As shown in Figure 3, if the reader has complete knowledge to understand the writer's document, the writer has only to provide the surface form of document. The reader's side of document architecture can recover every layer of document architecture by

using their own knowledge. However, if the reader has only partial knowledge, the author's side should compensate for the reader's knowledge.

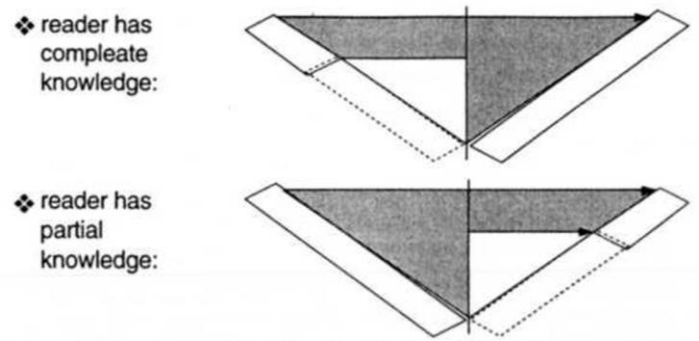


Figure 3. Cases of Information Architecture Network - Dynamics of Information

3.2. Configuration of Information Architecture Network and Normalization

Knowledge for recovering the complete document is interspersed with nodes in information architecture network as shown in Figure 4. When the knowledge is incomplete to recover the complete document, the reader (or user) questions the author's location and the author gives a solution. Such feedback processes and contents are logged in an interim node, which may be implemented as a logging server (physically in network). The information architecture network evolves after its self-organizing mechanism of learning and restructuring. The logging server is the source of such automatic evolution. It was claimed that the information architecture has processes between entities for information extraction and knowledge extraction. Because each process and entity is autonomous, nodes participated in the information architecture evolves autonomously. The process is not centralized but distributive. The participating node may not be one but redundant (not duplicated) for the same function. They are so compensating and competing each other.

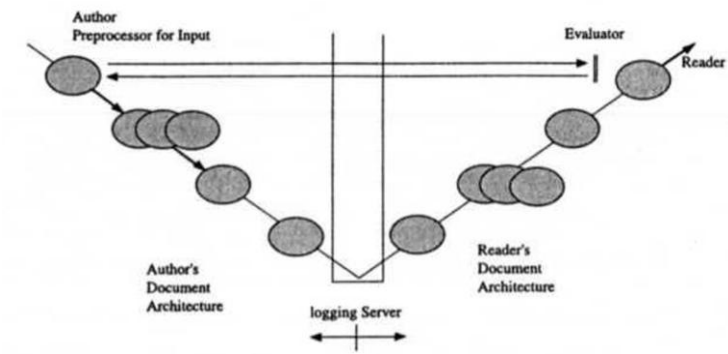


Figure 4. Information Architecture Network: Configuration

The information flow is incremental. At first, the surface form of document will be passed to the reader. When the reader cannot understand, it is rejected and returned to the author's location. At that time, the reader's side has an evaluator to measure whether the given form of document has enough information for the reader to understand or not.

In Figure 4, the document architecture is different and separate from each ocher. However, if there is a standardized form and process of document interchange format, they will be combined into one like Figure 5. The physical construction cost will be reduced and the operation will be more efficient.

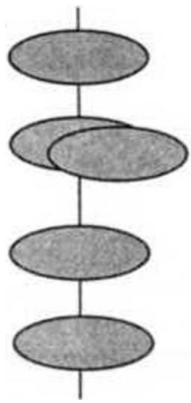


Figure 5. Distributed Information Architecture Network Based on Document Interchange Format Standards

4. Flexible Machine Translation for Multi-lingual Document Retrieval

4.1. Configuration of Flexible Machine Translation and Information Architecture

In fact, the process of translation does not assume understanding. In some occasion, a set of translation templates is enough to translate. For example, articles in stock news use only the special usage and special domain expression. That does not invoke the process to recover the complete document. If such process fails, then the next process in the deeper level of document architecture starts to be analyzed and translated. That process fills up the next level of document architecture, their result is transferred to the reader's language side by using the appropriate transfer knowledge. The sequence of processes to recover the complete document is awaken incrementally in a demand-based way. That is flexible in a sense that the process evocation is flexible. We call this "Flexible Machine Translation" (FMT). As shown in figure 6, after failure from the first process "morphology analyzer," that process either suggest the alternative solution or passes to the next process "syntactic analyzer". The result of syntactic analyzer is transferred based on syntactic pattern transfer knowledge. The evaluator of the corresponding node of syntactic generation in the reader's side measures whether the output of syntactic transfer is possible to be generated up to the final surface form. These processes continue whenever the reader's side sends the rejection signal. The feedback information is also logged just like the information architecture network (figure 4). The process will progress until the interlingua

meets. However, if the reader's side accepts the transferred document, the translation ends without going finally in the direction toward the point of interlingua.

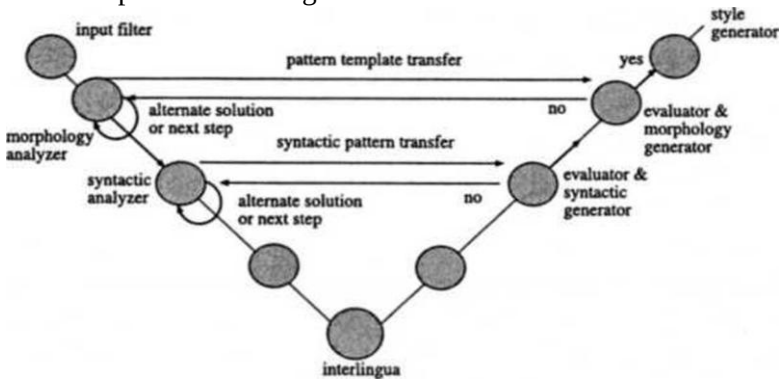


Figure 6. Flexible machine translation: Configuration

4.2. Distributed Flexible Machine Translation

As shown in the last section, every module in a flexible machine translation system can be linked to the (reader's) target language generation. Because every feedback is logged and stored onto the interim node, the system grows. Flexible machine translation systems are those of evolutionary and self-organizing networks.

Flexible machine translation is embodied in a "distributed" way. Every module can be a node in network. They are fault-tolerant because each level has competing nodes of the so-claimed same function. Every node is also competing and compensating processes or entities as shown in figure 7.

When we develop the flexible machine translation paradigm, the full system can operate from the beginning stage. The transfer of knowledge is a kind of document. The (author's) source sentence is a query to seek its component-wise patterns to be linked (or transferred) to the (reader's) target sentences. Such processes are the same as the processes of those of document retrieval.

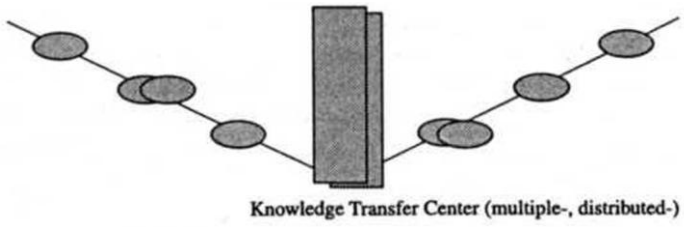


Figure 7. DFMT: Distributed Flexible Machine Translation Operation

5. Conclusion：Toward the Service of Intelligent Distributed Multi-lingual Document Retrieval

The balanced information flow is embodied in the document architecture and its standardization of document interchange format. The overall picture is drawn under the paradigm of information architecture. The mono-lingual document retrieval and the multi-lingual translation service is in one concept of information architecture. The “intelligence” of document retrieval assumes the full understanding, and that is shown to be a complete document.

In an operation of this paradigm, the standardization issue is one of practical objects for the successful communication. Assuming such standards, the knowledge to recover the complete document can be located in a “distributed”, network regardless of whether those are embodied virtually or physically. The practical cooperation in a distributed environment is possible under this paradigm. The operation starts from the beginning stage of development. These practical points support this paradigm: “document architecture under information architecture network” and its application: “intelligent distributed multilingual document retrieval".

References

Choi, Yong-Seok, Juho Lee, Jin-Xia Huang, and Key-sun Choi. 2000. *Cross-language information Retrieval System for Korean-Chinese-Japanese-English Languages*, Proceedings of ACL2000 (Demonstration), Hong Kong.

Russell, Stuart J. and Peter Norvig. 1994. "Artificial Intelligence: A Modem Approach", Prentice Hall.