# Interactive Paraphrasing Based on Linguistic Annotation

**Ryuichiro Higashinaka**
Keio Research Institute at SFC
5322 Endo, Fujisawa-shi,
Kanagawa 252-8520, Japan
rh@sfc.keio.ac.jp

**Katashi Nagao**
Dept. of Information Engineering
Nagoya University
Furo-cho, Chikusa-ku,
Nagoya 464-8603, Japan
nagao@nuie.nagoya-u.ac.jp

## Abstract

We propose a method “Interactive Paraphrasing” which enables users to interactively paraphrase words in a document by their definitions, making use of syntactic annotation and word sense annotation. Syntactic annotation is used for managing smooth integration of word sense definitions into the original document, and word sense annotation for retrieving the correct word sense definition for a word in a document. In this way, documents can be paraphrased so that they fit into the original context, preserving the semantics and improving the readability at the same time. No extra layer (window) is necessary for showing the word sense definition as in conventional methods, and other natural language processing techniques such as summarization, translation, and voice synthesis can be easily applied to the results.

## 1 Introduction

There is a large number of documents of great diversity on the Web, which makes some of the documents difficult to understand due to viewers’ lack of background knowledge. In particular, if technical terms or jargon are contained in the document, viewers who are unfamiliar with them might not understand their correct meanings.

When we encounter unknown words in a document, for example scientific terms or proper nouns, we usually look them up in dictionaries or ask experts or friends for their meanings. However, if there are lots of unfamiliar words in a document or there are no experts around, the work of looking the words up can be very time consuming. To facilitate the effort, we need (1) machine understandable online dictionaries, (2) automated consultation of these dictionaries, and (3) effective methods to show the lookup results.

There is an application which consults online dictionaries when the user clicks on a certain word on a Web page, then shows the lookup results in a popped up window. In this case, the application accesses its inner/online dictionaries and the consultation process is automated using the viewer’s mouse click as a cue. Popup windows correspond to the display method. Other related applications operate more or less in the same way.

We encounter three big problems with the conventional method.

First, due to the difficulty of word sense disambiguation, in the case of polysemic words, applications to date show all possible word sense candidates for certain words, which forces the viewer to choose the correct meaning.

Second, the popup window showing the lookup results hides the area near the clicked word, so that the user tends to lose the context and has to reread the original document.

Third, since the document and the dictionary lookup results are shown in different layers (e.g., windows), other natural language processing techniques such as summarization, translation, and voice synthesis cannot be easily applied to the results.

To cope with these problems, we realized a systematic method to annotate words in a document with word senses in such a way that anyone (e.g., the author) can easily add word sense information to a certain word using a user-friendly annotating tool. This operation can be considered as a creation of a link between a word in the document and a node in a domain-specific ontology.

The “Interactive Paraphrasing” that we propose makes use of word sense annotation and paraphrases words by embedding their word sense definitions into the original document to generate a new document.

Embedding occurs at the user’s initiative, which means that the user decides when and where to embed the definition. The generated document can also be the target for another embedding operation which can be iterated until the document is understandable enough for the user.

One of the examples of embedding a document into another document is *quotation.*

Transcopyright (Nelson, 1997) proposes a way for quoting hypertext documents.

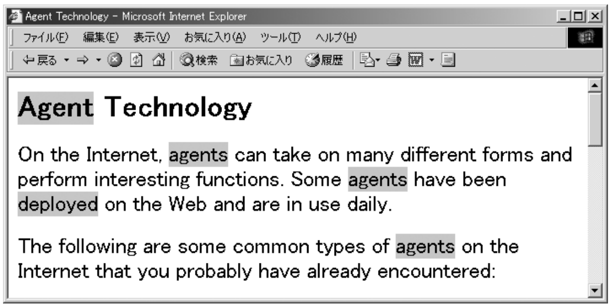
However, quoting means importing other documents as they are. Our approach is to convert other documents so that they fit into the original context, preserving the semantics and improving the readability at the same time.

As the result of embedding, there are no windows hiding any part of the original text, which makes the context easy to follow, and the new document is ready to be used for further natural language processing.

## 2 Example

In this section, we present how our system performs using screenshots.

Figure 1 shows an example of a Web document [1] after the automatic lookup of dictionary. Words marked with a different remains background color have been successfully looked up.


Agent Technology - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
戻る 検索 お気に入り 履歴

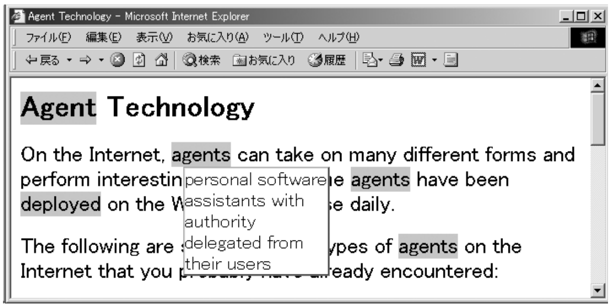
Agent Technology

On the Internet, agents can take on many different forms and perform interesting functions. Some agents have been deployed on the Web and are in use daily.

The following are some common types of agents on the Internet that you probably have already encountered:

Figure 1: Example of a web document showing dictionary lookup results

The conventional method such as showing the definition of a word in a popup window hides the neighboring text. (Figure 2)


Agent Technology - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
戻る 検索 お気に入り 履歴

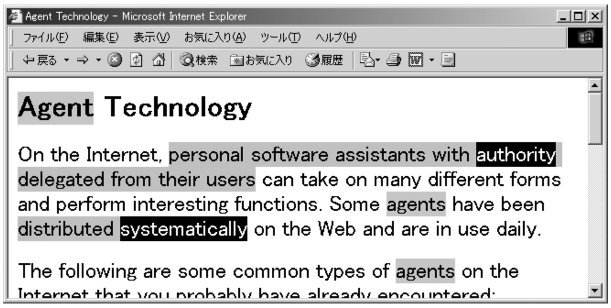
Agent Technology

On the Internet, agents can take on many different forms and perform interestin personal software assistants with authority delegated from their users e agents have been deployed on the W e daily.

The following are ypes of agents on the Internet that you eady encountered:

Figure 2: Example of a conventional method popup window for showing the definition

Figure 3 shows the result of paraphrasing the word "agent." It was successfully paraphrased using its definition "personal software assistants with authority delegated from their users." The word "deployed" was also paraphrased by the definition "to distribute systematically." The paraphrased area is marked by a different background color.


Agent Technology - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
戻る 検索 お気に入り 履歴

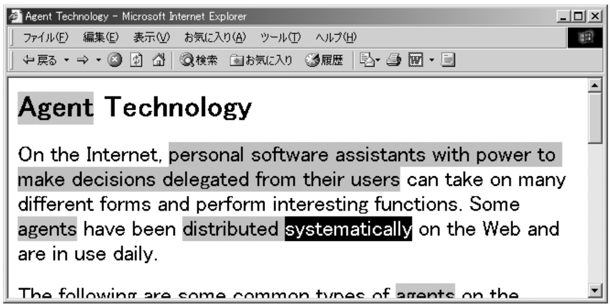
Agent Technology

On the Internet, personal software assistants with authority delegated from their users can take on many different forms and perform interesting functions. Some agents have been distributed systematically on the Web and are in use daily.

The following are some common types of agents on the Internet that you probably have already encountered:

Figure 3: Example of the results after paraphrasing "agents" and "deployed"

Figure 4 shows the result of paraphrasing the word in the area already paraphrased. The word "authority" was paraphrased by its definition "power to make decisions."


Agent Technology - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
戻る 検索 お気に入り 履歴

Agent Technology

On the Internet, personal software assistants with power to make decisions delegated from their users can take on many different forms and perform interesting functions. Some agents have been distributed systematically on the Web and are in use daily.

The following are some common types of agents on the

Figure 4: Example of incremental paraphrasing

## 3 Linguistic Annotation

Semantically embedding word sense definitions into the original document without changing the original context is much more difficult than showing the definition in popup windows.

For example, replacing some word in a sentence only with its word sense definition may cause the original sentence to be grammatically wrong or less cohesive.

This is due to the fact that the word sense definitions are usually incapable of simply replacing original words because of their fixed forms.

For appropriately integrating the word sense definition into the original context, we employ syntactic annotation (described in the next section) to both original documents and the word

[1] This text, slightly modified here, is from "Internet Agents: Spiders, Wanderers, Brokers, and Bots," Fah-Chun Cheong, New Riders Publishing, 1996.

sense definitions to let the machine know their contexts.

Thus, we need two types of annotations for Interactive Paraphrasing. One is the word sense annotation to retrieve the correct word sense definition for a particular word, and the other is the syntactic annotation for managing smooth integration of word sense definitions into the original document.

In this paper, linguistic annotation covers syntactic annotation and word sense annotation.

### 3.1 Syntactic Annotation

Syntactic annotation is very useful to make online documents more machine-understandable on the basis of a new tag set, and to develop content-based presentation, retrieval, question-answering, summarization, and translation systems with much higher quality than is currently available. The new tag set was proposed by the GDA (Global Document Annotation) project (Hasida, *http://www.etl.go.jp/etl/nl/gda/*). It is based on XML , and designed to be as compatible as possible with TEI (The Text Encoding Initiative, *http://www.uic.edu:80/orgs/tei/*) and CES (Corpus Encoding Standard, *http://www.cs.vassar.edu/CES/*). It specifies modifier-modifiee relations, anaphor-referent relations, etc.

An example of a GDA-tagged sentence is as follows:

```
<su><np rel="agt">Time</np>
<v>flies</v><adp rel="eg">
<ad>like</ad><np>an <n>arrow</n></np>
</adp>.</su>
```

The tag, `<su>`, refers to a sentential unit. The other tags above, `<n>`, `<np>`, `<v>`, `<ad>` and `<adp>` mean noun, noun phrase, verb, adnoun or adverb (including preposition and postposition), and adnominal or adverbial phrase, respectively.

Syntactic annotation is generated by automatic morphological analysis and interactive sentence parsing.

Some research issues concerning syntactic annotation are related to how the annotation cost can be reduced within some feasible levels. We have been developing some machine-guided annotation interfaces that conceal the complexity of annotation. Machine learning mechanisms also contribute to reducing the cost because they can gradually increase the accuracy of automatic annotation.

### 3.2 Word Sense Annotation

In the computational linguistic field, word sense disambiguation has been one of the biggest issues. For example, to have a better translation of documents, disambiguation of certain polysemic words is essential. Even if an estimation of the word sense is achieved to some extent, incorrect interpretation of certain words can lead to irreparable misunderstanding.

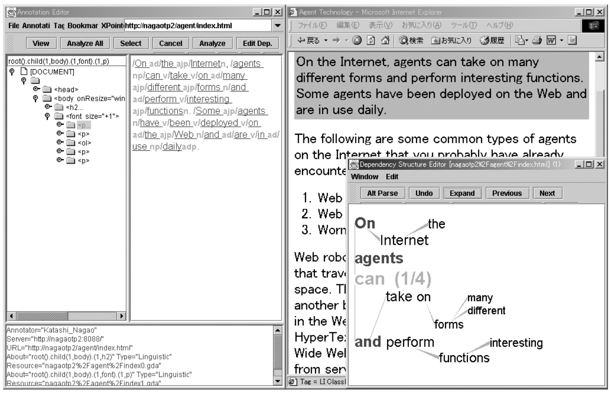
To avoid this problem, we have been promoting annotation of word sense for polysemic words in the document, so that their word senses can be machine-understandable.

For this purpose, we need a dictionary of concepts, for which we use existing domain ontologies. An ontology is a set of descriptions of concepts - such as things, events, and relations - that are specified in some way (such as specific natural language) in order to create an agreed-upon vocabulary for exchanging information.

Annotating a word sense is therefore equal to creating a link between a word in the document and a concept in a certain domain ontology. We have made a word sense annotating tool for this purpose which has been integrated with the annotation editor described in the next section.

### 3.3 Annotation Editor

Our annotation editor, implemented as a Java application, facilitates linguistic annotation of the document. An example screen of our annotation editor is shown in Figure 5.



/On ad/the ajp/Internet n, /agents np/can v/take v/on ad/many ajp/different ajp/forms n/and ad/perform v/interesting ajp/functions n. /Some ajp/agents n/have v/been v/deployed v/on ad/the ajp/Web n/and ad/are v/in ad/use np/daily adp.

Annotator="Katashi_Nagao"
Server="http://nagaotp2:8088/"
URL="http://nagaotp2/agent/index.html"
About="root().child(1,body).(1,h2)" Type="Linguistic"
Resource="nagaotp2%2Fagent%2Findex0.gda"
About="root().child(1,body).(1,font).(1,p)" Type="Linguistic"
Resource="nagaotp2%2Fagent%2Findex1.gda"

On the Internet, agents can take on many different forms and perform interesting functions. Some agents have been deployed on the Web and are in use daily.

The following are some common types of agents on the Internet that you probably have already encounte

1. Web
2. Web
3. Worn

Web robo that trav space. Tl another l in the We HyperTe: Wide Wel from ser

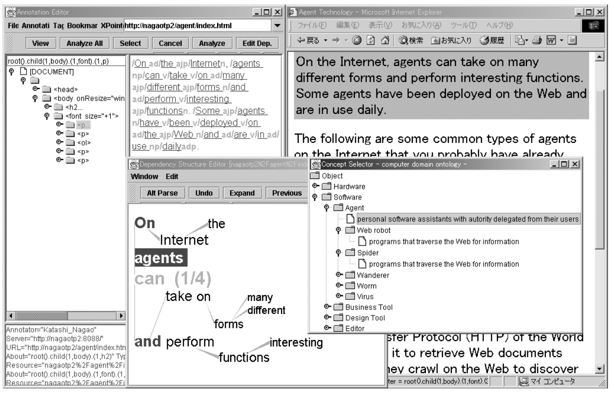
Figure 5: Annotation editor

The left window of the editor shows the document object structure of the HTML document. The center window shows some text that was selected on the Web browser as shown on the right top of the figure. The selected area is automatically assigned an XPointer (i.e., a location identifier in the document) (World Wide Web Consortium, *http://www.w3.org/TR/xptr/*).

The right bottom window shows the linguistic structure of the sentence in the selected area. In this window, the user can modify the results of the automatically-analyzed sentence structure.

Using the editor, the user annotates text with linguistic structure (syntactic and semantic structure) and adds a comment to an element in the document. The editor is capable of basic natural language processing and interactive disambiguation.

The tool also supports word sense annotation as shown in Figure 6. The ontology viewer appears in the right middle of the figure. The user can easily select a concept in the domain ontology and assign a concept ID to a word in the document as a word sense.



/On ad/the ajp/Internetn, /agents np/can v/take v/on ad/many ajp/different ajp/forms n/and ad/perform v/interesting ajp/functionsn. /Some ajp/agents n/have v/been v/deployed v/on ad/the ajp/Web n/and ad/are v/in ad/use np/dailyadp.

Annotator="Katashi_Nagao"
Server="http://nagaotp2:8088/"
URL="http://nagaotp2/agent/index.htm
About="root().child(1,body).(1,h2)" Typ
Resource="nagaotp2%2Fagent%2Fi
About="root().child(1,body).(1,font).(1,
Resource="nagaotp2%2Fagent%2Fi

On the Internet, agents can take on many different forms and perform interesting functions. Some agents have been deployed on the Web and are in use daily.

The following are some common types of agents on the Internet that you probably have already

sfer Protocol (HTTP) of the World
it to retrieve Web documents
ey crawl on the Web to discover

Annotation Editor
File Annotati Tag Bookmar XPoint http://nagaotp2/agent/index.html
View
Analyze All
Select
Cancel
Analyze
Edit Dep.
root().child(1,body).(1,font).(1,p)
[DOCUMENT]
<head>
<body onResize="win
<h2...
<font size="+1">
<p...
<p>
<ol>
<p>
<p>
Agent Technology - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
戻る
検索
お気に入り
履歴
Dependency Structure Editor [nagaotp2%2Fagent%2Find
Window Edit
Alt Parse
Undo
Expand
Previous
On
the
Internet
agents
can (1/4)
take on
many
different
forms
and perform
interesting
functions
Concept Selector - computer domain ontology -
Object
Hardware
Software
Agent
personal software assistants with autority delegated from their users
Web robot
programs that traverse the Web for information
Spider
programs that traverse the Web for information
Wanderer
Worm
Virus
Business Tool
Design Tool
Editor
ter = root().child(1,body).(1,font).(
マイ コンピュータ

Figure 6: Annotation editor with ontology viewer

## 4 Interactive Paraphrasing

Using the linguistic annotation (syntactic and word sense annotation), Interactive Paraphrasing offers a way to paraphrase words in the document on user demand.

### 4.1 Interactivity

One of the objectives of this research is to make online documents more understandable by paraphrasing unknown words using their word sense definitions.

Users can interactively select words to paraphrase by casual movements like mouse clicks.

The paraphrase history is stored for later use such as profile-based paraphrasing (yet to be developped) which automatically selects words to paraphrase based on user's knowledge.

The resulting sentence can also be a target for the next paraphrase. By allowing incremental operation, users can interact with the document until there are no paraphrasable words in the document or the document has become understandable enough.

Interactive Paraphrasing is divided into *click paraphrasing* and *region paraphrasing* according to user interaction type. The former paraphrases a single word specified by mouse click, and the latter, one or more paraphrasable words in a specified region.

### 4.2 Paraphrasing Mechanism

As described in previous sections, the original document and the word sense definitions are annotated with linguistic annotation, which means they have graph structures. A word corresponds to a node, a phrase or sentence to a subgraph. Our paraphrasing is an operation that replaces a node with a subgraph to create a new graph. Linguistic operations are necessary for creating a graph that correctly fits the original context.

We have made some simple rules (principles) for replacing a node in the original document with a node representing the word sense definition.

There are two types of rules for paraphrasing. One is a "global rule" which can be applied to any pair of nodes, the other is a "local rule" which takes syntactic features into account.

Below is the description of paraphrasing rules (principles) that we used this time. $Org$ stands for the node in the original document to be paraphrased by $Def$ which represents the word sense definition node. Global rules are applied first followed by local rules. Pairs to which rules cannot be applied are left as they are.

**- Global Rules -**

1. If the word $Org$ is included in $Def$, paraphrasing is not performed to avoid the loop of $Org$.
2. Ignore the area enclosed in parentheses in $Def$. The area is usually used for making $Def$ an independent statement.
3. Avoid double negation, which increases the complexity of the sentence.
4. To avoid redundancy, remove from $Def$ the same case-marked structure found both in $Org$ and $Def$.
5. Other phrases expressing contexts in $Def$ are ignored, since similar contexts are likely to be in the original sentence already.

**- Local Rules -**

The left column shows the pair of linguistic features [2] corresponding to $Org$ and $Def$. (e.g. $N - N$ signifies the rule to be applied between nodes having noun features.)

[2] $N$ stands for the noun feature, $V$, $AJ$ and $AD$ for verbal, adjective and adverbial features respectively.

| | |
|---|---|
| $N-N$ | Replace $Org$ with $Def$ agreeing in number. |
| $N-V$ | Nominalize $Def$ and replace $Org$. (e.g., explain → the explanation of) |
| $V-N$ | If there is a verbal phrase modifying $Def$, conjugate $Org$ using $Def$'s conjugation and replace $Org$. |
| $V-V$ | Apply $Org$'s conjugation to $Def$ and replace $Org$. |
| $AD-N$ | Replace $Org$ with any adverbial phrase modifying $Def$. |
| $AJ-N$ | Replace $Org$ with any adjective phrase modifying $Def$. |

### 4.3 Implementation

We have implemented a system to realize Interactive Paraphrasing. Figure 7 shows the basic layout of the system. The proxy server in the middle deals with user interactions, document retrievals, and the consultation of online dictionaries.

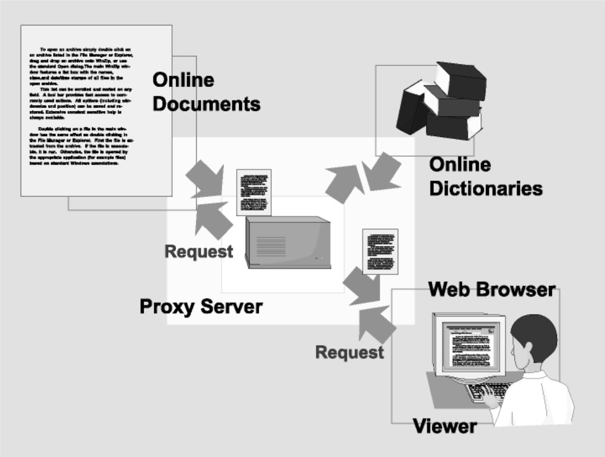


Figure 7: System architecture

The paraphrasing process follows the steps described below.

1. On a user's request, the proxy server retrieves a document through which it searches for words with word sense annotations. If found, the proxy server changes their background color to notify the user of the paraphrasable words.
2. The user specifies a word in the document on the browser.
3. Receiving the word to be paraphrased, the proxy server looks it up in online dictionaries using the concept ID assigned to the word.
4. Using the retrieved word sense definition, the proxy server attempts to integrate it into the original document using linguistic annotation attached to both the definition and the original document.

## 5 Related Work

Recently there have been some activities to add semantics to the Web (Nagao et al., 2001) (SemanticWeb.org, *http://www.semanticweb.org/*) (Heflin and Hendler, 2000) enabling computers to better handle online documents. As for paraphrasing rules concerning structured data, Inui et al. are developing Kura (Inui et al., 2001) which is a Transfer-Based Lexico-Structural Paraphrasing Engine.

## 6 Conclusion and Future Plans

We have described a method, "Interactive Paraphrasing", which enables users to interactively paraphrase words in a document by their definitions, making use of syntactic annotation and word sense annotation.

By paraphrasing, no extra layer (window) is necessary for showing the word sense definition as in conventional methods, and other natural language processing techniques such as summarization, translation, and voice synthesis can be easily applied to the results.

Our future plans include: *reduction of the annotation cost, realization of profile-based paraphrasing using personal paraphrasing history, and retrieval of similar pages for semantically merging them using linguistic annotation.*

## References

Jeff Heflin and James Hendler. 2000. Semantic Interoperability on the Web. In *Proceedings of Extreme Markup Languages 2000. Graphic Communications Association, 2000. pp. 111-120.*

Kentaro Inui, Tetsuro Takahashi, Tomoya Iwakura, Ryu Iida, and Atsushi Fujita. 2001. KURA: A Transfer-Based Lexico-Structural Paraphrasing Engine. In *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium, Workshop on Automatic Paraphrasing: Theories and Applications.*

Katashi Nagao, Yoshinari Shirai, and Kevin Squire. 2001. Semantic annotation and transcoding: Making Web content more accessible. *IEEE MultiMedia. Vol. 8, No. 2, pp. 69–81.*

Theodor Holm Nelson. 1997. Transcopyright: Dealing with the Dilemma of Digital Copyright. *Educom Review, Vol. 32, No. 1, pp. 32-35.*