An Annotation Tool for Multimodal Dialogue Corpora using Global Document Annotation

Kazunari ITO and Hiroaki SAITO
Keio University
Department of Science for Open and Environmental Systems
3-14-1 Hiyoshi, Kohoku-ku, Yokohama, Japan, 223-8522
{k_ito , hxs}@nak.ics.keio.ac.jp

Abstract

This paper reports a tool which assists the user in annotating a video corpus and enables the user to search for a semantic or pragmatic structure in a GDA tagged corpus. An XQL format is allowed for search patterns as well as a plain phrase. This tool is capable of generating a GDA time-stamped corpus from a video file manually. It will be publicly available for academic purposes.

1 Introduction

To achieve a natural communication environment between computers and the users, many interactive prototype systems that can talk with the user have been developed using multimodal information (face expressions, voice tones, gestures, etc.). Since multimodalness of these systems is manually built in, achieving free and effective communication or enhancing communication ablilities is not easy. Thus automatic learning from huge data is hoped for.

Recently such various video data as TV dramas, news, and language teaching materials are available, from which natural interactiveness should be extracted. Such interactiveness from an intellectual content is also valuable for the fields of machine translation, information retrieval, handling question responses, and knowledge discovery systems.

GDA[1] (Global Document Annotation), which is an XML tag set, adds information on syntax, semantics, and pragmatics to texts (Hashida 1998). The texts with GDA organically corresponding to voice and a video will contribute to the basic research into these technologies and promote the application development.

2 GDA tagged corpus

This chapter explains the GDA tag set and a method which relates tagged data with the video image.

2.1 GDA

The GDA Initiative aims at having Internet authors annotate their electronic documents with a common standard tag set which allows machines to automatically recognize the semantic and pragmatic structures of the documents. A huge amount of annotated data is expected to emerge, which should serve not just as tagged linguistic corpora but also as a worldwide, self-extending knowledge-base mainly consisting of examples of how our knowledge manifests. It describes the meaning of sentence analysis (semantics and pragmatics) basically. It also describes information on the subject role, the rhetoric relation, and correspondence. Figure 1 shows an example of the text “みみはかくれてる (an ear is covered)” tagged with GDA. Note that GDA is totally language independent, although all the following examples include Japanese texts.

[1] http://www.i-content.org/GDA/tagset.html

```
<q who="A">
  <su syn=″fc″ id="kakure">
     <vp syn="f">
       <n arg="X">みみ</n>
       <ad sem="obj">は</ad>
       <adp syn="f">
         <v>かくれ</v>
         <ad>て</ad>
       </adp>
       <v>る</v>
     </vp>
   </su>
</q>
```

Figure 1: A fragment of GDA corpus

In Figure 1, <q> represents a word where that part is spoken by someone, and the value of who attribute shows who utters it. <su> indicates a sentence, having no syntactic relation to other parts of the utterance. Attribute "syn" means it is a syntax structure of the sentence, and "fc" means a forward link dependency. <v> and <vp> elements mean a verb and a verb phrase, respectively. <n> element represents a noun or a noun phrase. <ad> and <adp> elements are an adverb, and a postpositional phrase. As these examples display, the GDA tag set has been determined to show syntactic structure effectively where a word is assumed to be a unit.

2.2 Adding Time -information to GDA tagged corpora

When you relate the video image file with its text file, it is widely used to embed the time information indicating when an utterance is spoken. We define the following two kinds of new formats to relate a video file with its GDA corpus file .

1. btime and etime attributes: These attributes can be added to an arbitrary tag. Attribute btime shows the start time of an utterance. Attribute etime shows the finished time. The format is described as follows.

```
<any btime=″utterance start time(sec)″
     etime=″utterance end time(sec)″ >
     sentence
</any>
```

With these attributes, the voiceless section of an utterance can be precisely indicated and the overlap event is detected in a multi-speaker environment.

2. tst (time stamp) tag: Tag "tst" is an empty element tag and described in the following format.

```
<tst val = ″utterance start time
(sec)″ />
```

A tst tag can be inserted in an arbitrary place. The ending time of an utterance can be determined from the value of the next btime attribute or val attribute of the next tst tag. Moreover, it is also possible that the tst tag has btime and etime attribute. Figure 2 shows an example when time information is added to the GDA tagged corpus in Figure 1. You can see a man allocated "A" speaks a word "みみ (an ear)" from 60.81 to 61.03 sec, a word "は (is)" from 61.10 to 61.90 sec.

```
<q who=″A″>
  <su id=″kakure″>
    <vp syn=″f″>
      <n arg="X" btime=″60.815″
                etime="61.034″>みみ
       </n>
       <ad sem=″obj″>
        は<tst val=″61.100″/></ad>
       <adp syn=″f″>
         <v>かくれ
         <tst val="61.907"/></v>
         <ad>て<tst val="62.192"/></ad>
       </adp>
       <v>る<tst val="62.383"/></v>
     </vp>
   </su>
</q>
```

Figure 2: An example of GDA corpus with timestamp

These time annotation is often inserted after GDA tagging, because time information is independent from the standard syntactic/pragmatic GDA.

3 An annotation tool for multimodal dialogue corpus

We have developed a multimodal annotation tool for a video corpus and its annotated data (JEITA 2001). This tool runs on any platform that accommodates Java2 and Java Media Framework (JMF[2]) ver.2.0 or higher. It also requires Java-based XML parser Xerces-J [3] and Java-based XQL engine

[2] http://java.sun.com/products/java-media/jmf/

[3] http://xml.apache.org/xerces-j/

GMD-IPSI XQL[4]. Moreover, users can easily extend functions by mounting plug-ins. The prototype is equipped with two different types of plug-ins. One is "XQL search plug-in" in which the user can search for various syntactic and semantic patterns in a GDA corpus. An XQL format query as well as a plain word is allowed for search patterns. Another is "Annotation plug-in" which assists the annotator in generating a time-stamped GDA corpus from a video file.

4 Basic Functions

A screenshot of basic composition is shown in Figure 3. The whole window is composed of several internal windows.

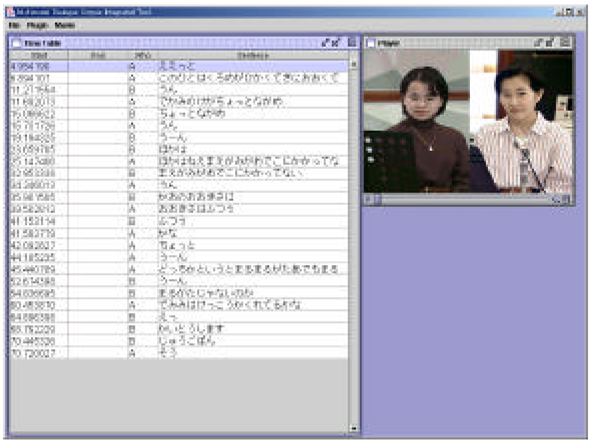


Figure 3: Screenshot of basic composition

The left internal window is "time table window" which displays each sentence in a table format. When the user clicks one of the columns in the table, the video corresponding to the sentence is played. Rows of the table are highlighted consecutively during that part of the video is being displayed. The top right is "video image window". It displays and plays the video image.

5 Extended Functions

This section explains the outline of two plug-ins first and usage of these after that.

5.1 XQL search plug-in

Figure 4 shows the screenshot when the XQL search plug-in is loaded, you can see new window appears at the bottom left of the tool. The user can type a query into the text field at the bottom of the new window.

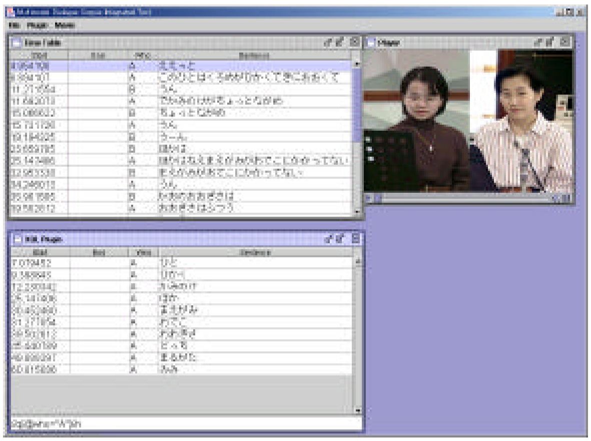
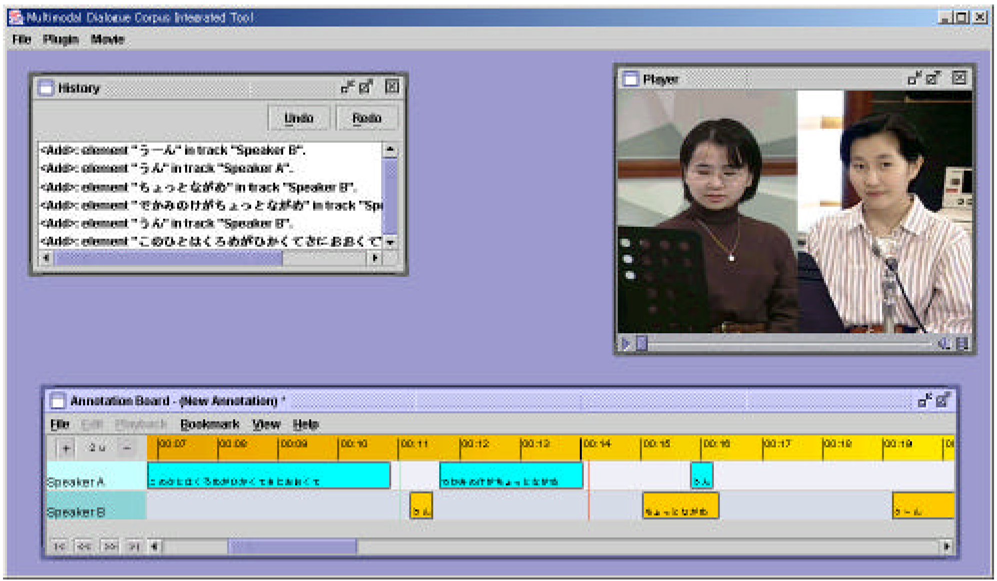


Figure 4: Screenshot when XQL search plug-in is loaded

An acceptable query format is a plain text or a query text defined by XQL (XML Query Language). XQL is a subset of query language XQuery that uses XML as a data model, a recommendation by W3C. XQL has already been mounted on software over many fields like Web browsers, document archiving systems, XML middleware, Perl libraries, and a command line utility. XQL always returns a part of the document. In XQL, the hierarchy of the node is written by '/', an arbitrary hierarchy is written by '//', the attribute name by '@name', the tag name by 'name' as it is. A regular expression and a conditional expression are also acceptable. For a exmple, ' //q[@who='A']' returns 'q' node with the value of attribute name 'who' equals 'A'.

Figure 5 shows search results. In Figure 5, five words [かみのけ(a hair), まえがみ(a forelock), おでこ(a prominent forehead), かお(a face), みみ(an ear)] are matched on the condition that the part of speech is noun and the value of the attribute "arg" equals to "X" (see the fourth line of Figure 2). Moreover, since the label "X" is attached to the utterance "この人(this person)" in this GDA file, this query becomes a meaning of extracting a noun phrase whose subject is "this person". The user can search for the subject even if an object is omitted in suited sentences.

[4] http://xml.darmstadt.gmd.de/xql/

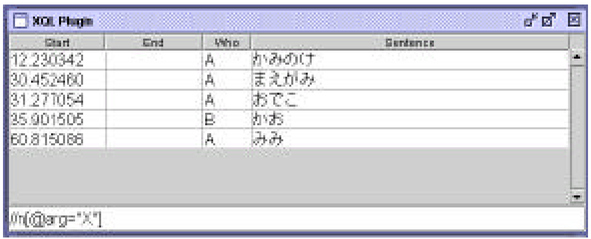


Multimodal Dialogue Corpus Integrated Tool
File Plugin Movie
History
Undo
Redo
Player
Annotation Board - (New Annotation) *
File Edit Playback Bookmark View Help
Speaker A
Speaker B

<Add>: element "う－ん" in track "Speaker B".
<Add>: element "うん" in track "Speaker A".
<Add>: element "ちょっとながめ" in track "Speaker B".
<Add>: element "でかみのけがちょっとながめ" in track "Sp
<Add>: element "うん" in track "Speaker B".
<Add>: element "このひとはくろめがひかくてきにおおくて

Figure 6: Screenshot when Annoation plug-in is loaded

A query of extracting a synonym of a certain word, only the utterance of a specific speaker to another, and a sentence that of the response for a certain utterance, etc. are possible. Semantic or discourse structures are also extractable if such information is annotated in the file. Clicking any column of the table, the corresponding media section is played like the "time table Window".



Start	End	Who	Sentence
12.230342		A	かみのけ
30.452460		A	まえがみ
31.271054		A	おでこ
35.901505		B	かお
60.815086		A	みみ

//n[@arg="X"]

Figure 5: Result by query [a noun and value of attribute "arg" equals "X"]

5.2 Annotation plug-in

Creating a GDA file with timestamps record automatically with less time and less labor is indispensable to make large quantities of them and to spread them. From the accuracy of the current speech recognition technology, it is difficult to attach timestamps record automatically and accurately by taking synchronization of a video image. Annotation plug-in increases the efficiency of time-stamping a GDA file by visual operation. Figure 6 shows the screenshot when it is loaded.

The window located on the lower part of Figure 6 called "Annotation board" which displays information on a GDA file with a time-stamp visually. You can also see a horizontal axis which expresses the time on media, and two layers in the board. Upper layer displays utterances of "Speaker A", lower layer displays "Speaker B" in this case. Rectangles on one layer represent each speaker's utterances according to the time series. The utterance itseft is displayed in the rectangle, color of which is different for each speaker. Length and the position indicate time information of the utterance. A person edits the annotation by operating two kinds of lines on the board. "Current line" shows a current playback position on the media. "Base line" indicates the start or the end point of the time-stamp when the utterance sentence is newly inserted. Various functions (change of line's position, and deletion or insertion of an utterance) can be executed by mouse operation. When the annotator clicks on the board, the "Current line" moves and a frame to which a video image corresponds is displayed. Thus, an annotator can attach the information of the start time and the end time and utterance text itself manually. A prototype system which automatically converts the GDA file from raw text files with a morphological analysis and a parsing tool in addition to the original filter has already been proposed (Suzuki 2001).

6 Current development

The core functions of the tool are complete and stable. Still, there is much room for expansion. The following functions are being developed.

6.1 User-friendly GUI-based search interface

Needless to say, there is no guarantee that aclause or a sentence which agrees with an XQL query exists. Moreover the user has to know the XQL expression for search. We believe it necessary that a retrieval way by the top-down philosophy which narrows the candidates while presenting suited clauses sequentially.

It is very difficult at present for an XQL to express dependency relations among the search condition. Now, a query language of XML has been integrated into Xquery. Hence, we are scheduled to bulid an Xquery engine. A user-friendly GUI-based search window for the retrieval which does not require an explicit XQL query is currently being developed.

6.2 Uniting with other multimodal data

There are many kinds of specifications to describe multimodal data. For example, J-ToBI(Venditti, 1995) which describes prosodic information of voice, FACS(Ekman, etc. 1978) which annotates person's expression, etc. We are scheduled to design the specification to integrate these information into GDA in the XML format. As a result, the user will be able to present a condition, for example, of a word ‘Truth’ of doubt type or ‘I see’ of hesitation.

It is also scheduled to relate visual information of video data with GDA. They can contain information on motion, glance and the place of the object in video image. A reverse-search which extracts a corresponding text from visual information in a video image becomes available, too.

6.3 Coordinated functions

We intend to enhance a relation with other annotation tools. Concretely, various formats of output files can be taken in this tool in XML formats. We shall define a DTD (data conversion definition) to enable export and import to/from other tools. A function of date exchange enables the user to enhance flexibility and accessbility of this tool.

6.4 Retrieval for multiple files

The user can retrieve only a single file in a local machine at present. This tool will cope with the client-server model that it requests retrieval demand to the corpora database servers on a network, downloads only necessary files to the local machine and analyzes them.

7 Related works

Most of recent multilmodal annotation tool projects are almost Java-based, use XML for file exchange and have an object-oriented design:

MATE (Carletta, etc. 1999) is an annotation workbench that allows highly generic specification via stylesheets that determine shape and functionality of the user’s implemation. Speed and stability for the tool are both still problematic for real annotation. Also, the generic approach increases the initial effort to set up the tool since you basically have to write your own tool using the MATE style language.

EUDICO (Brugman et al. 2000) is an effort to allow multi-user annotation of a centrally localted corpus via a web-based interface. The tools that are available to work with the multimodal corpus make it possible to analyze their content and to add new free defined annotations to them. A EUDICO client can choose a subset of the corpus data.

Anvil (Kipp, 2001) is a generic video annotation tool which allows frame-accurate, hierarchical multi-layered annotation with objects that contain attribute-value pairs. Layers and attributes are all user-defineable. A time-aligned view and some configuration options make coding work quite intuitive.

ATLAS project (Steven, etc. 2000) deals with all types of annotation and is theoretically based on the idea of annotation graphs where nodes represent time points and edges indicate annotation labels.

8 Conclusion

We have reported an annotation tool for multimodal dialogue corpora. This tool enables semantic and pragmatic search from a video data with annotated texts in the GDA format. This tool is platform-independent and equipped with easy-to-use interface. It will be of use to researchers dealing with multimodal dialogue for exploratory studies as well as annotation. Core functions are complete and various extension facilities are now being developed. This prototype will be publicly available soon.

9 Acknowledgements

We wish to express our gratitute for the members of the committee on methods and standards of dialogic content in Japan Electronics and Information Technology Industries Association (JEITA). Especially we would like to thank Koichi Hashida for insightful advices.

References

Hasida, K. (1998) Intellectual contents of all-round based on GDA meaning modification. The transaction of Japanese Society for Artificial Intelligence., Vol. 13, No.4, pp.528-535 (in Japanese).

JEITA (2001) Servey Report about natural language processing system. pp.49-56(in Japanese).

J. Suzuki and K. Hasida (2001) Proposal of answer extraction system using GDA tag. The 7th annual meeting of Language Processing Society (in Japanese).

Ekman, P. and Friesen, W.(1978) Facial action coding system : a technique for the measurement of facial-movement, Consulting Psychologists Press, 1978

J.J. Venditti,(1995) Japanese ToBI Labelling Guidelines. Ohio-State University,Columbus,U.S.A.,1995.

Michael Kipp (2001) Anvil - A Generic Annotation Tool for Multimodal Dialogue, Proceedings of Eurospeech 2001, pp.1367-1370.

Carletta, J. and Isard, A (1999) The MATE Annotation Workbench, In Proceedings of the ACL Workshop, Towards Standards and Tools for Discourse Tagging., pp.11-17.

H. Brugman, A. Russel, D. Broeder, and P.Wittenburg (2000) EUDICO. Annotation and Exploitation of Multi Media Corpora, Proceedings of LREC 2000 Workshop.

Steven Bird, David Day, John Garofolo, John Henderson, Christophe Laprun, and Mark Liberman (2000) ATLAS: A Flexible and Extensible Architecture for Linguistic Annotation, Proceedings of the Second International Conference on Language Resource and Evaluation, pp.1699-1706.