Machine Translation Based on NLG from XML-DB

Yohei Seki
Aoyama Gakuin / Department of Informatics,
University The Graduate University
for Advanced Studies
(Sokendai)

Ken'ichi Harada
Department of Computing Science
Keio University

Abstract

The purpose of this study is to propose a new method for machine translation. We have proceeded through with two projects for report generation (Kittredge and Polguere, 2000) : Weather Forecast and Monthly Economic Report to be produced in four languages : English, Japanese, French, and German. Their input data is stored in XML-DB. We applied a three-stage pipelined architecture (Reiter and Dale, 2000), and each stage was implemented as XML transformation processes. We regard XML stored data as language-neutral intermediate form and employ the so-called 'sublanguage approach' (Somers, 2000). The machine translation process is implemented via XML-DB as a kind of interlingua approach instead of the conventional structure transfer approach.

1 Introduction

As the variety of users accessing the common resources on the World Wide Web, the importance of multimedia and multilingual information presentation technology has increased. Machine translation technology is essential for multilingual presentations, and many researchers pursue language independent structures; i.e. Rassinoux et al. (1998), etc. Many semantic structures like 'semantic frame' or 'feature structures' have been developed and common language attributes were embedded in these structures.

On the other hand, to store the resources, there were lots of databases all over the world. Those DBs stored in relational style with numerical data format would necessarily have language independent features. There, however, was a gap between DB structures and language independent semantic structures. Recently, the XML-DB technologies have been developed to support a more structured function for databases. The structuring techniques are useful not only for data structure but to represent linguistic structures.

We developed the XML-based report generation system as in Figure 1. The system is based on a three-stage pipelined architecture : document planning, microplanning, and surface realization. This system produces four languages : Japanese, English, French, and German from the common resources. The system also supports the publishing in VoiceXML format[1] and synthesis function by using IBM websphere VoiceServer SDK[2]. Therefore, this system is also useful for handicapped people like the visually handicapped to share information.

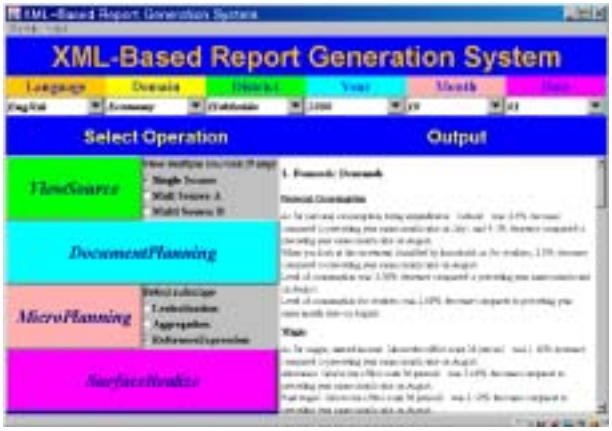

Figure 1. XML-Based Report Generation System

This paper consists of four sections. In Section 2, we discuss language independent and dependent features. Section 3 details multilingual generation and voice synthesis technologies in our system. Finally, in Section 4, we present our conclusions.

[1] http ://www.voicexml.org
[2] http ://www-3.ibm.com/software/speech/enterprise/ep_11.html

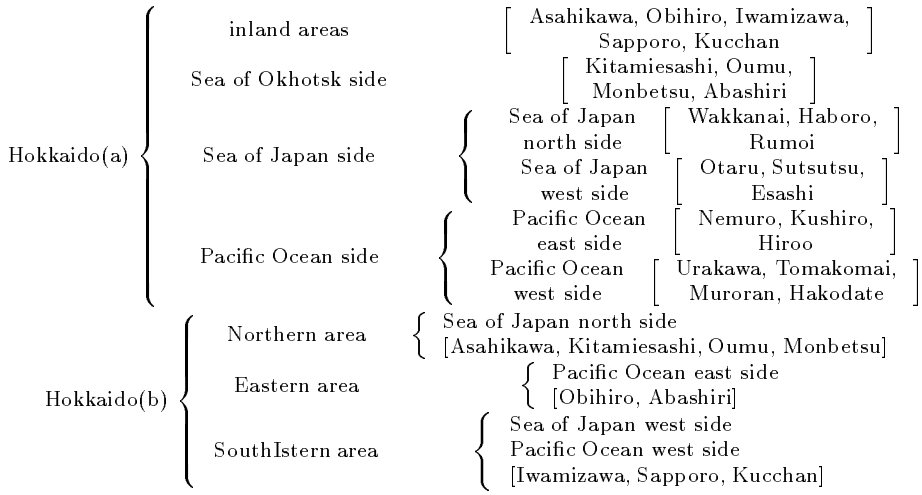

FIG. 2: The Ontologies with District Names for Weather Forecasts in Hokkaido

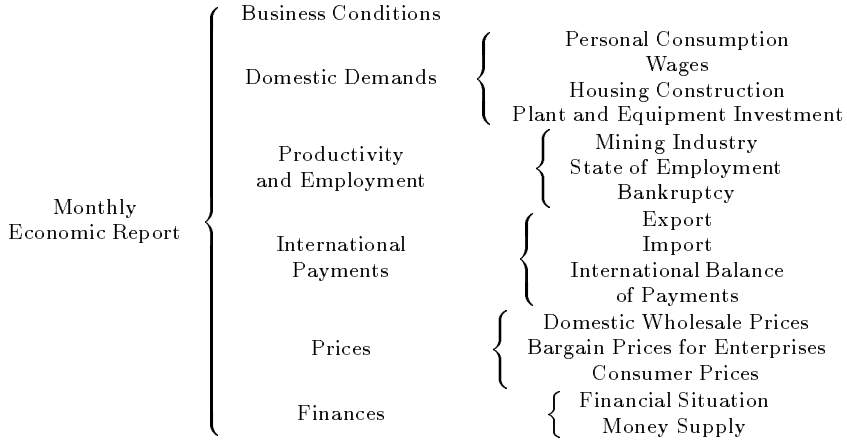

FIG. 3: The Ontologies for Monthly Economic Reports

2 Language Independence and Dependency on Discourse Unit

2.1 Language Independence

The intentional structure to retrieve data from DB is based on language-independent ontological structure (Rassinoux et al., 1998). Although we take the input database which is originally provided with RDB style format[3], the input data is restructured according to an ontologically-formulated structure with XML format. In case of the 'weather forecast', our data from each observatory point was structured according to place names. On the other hand, in case of the 'economy report', each numeric item was structured according to the contents of each heading. These structures are shown in Figures 2 and 3.

2.2 Lexical Variation Depends on Discourse Unit

In order to generate individual languages, lexical paraphrasing processes according to discourse units must be carried out. For example, the numerical data concerning 'increasing' and 'decreasing' expressions in Japanese varies according to the subject ; i.e. prices : increasing = 'joushou' ; decreasing = 'geraku' and other subjects : increasing = 'zou' ; decreasing = 'gen'. We implement this paraphrasing as a surface

[3]In fact, our weather domain input data is the Annual Report published by the Japan Meteorological Agency (http ://www.jmbsc.or.jp/offline/cd0040.htm). On the other hand, our economy domain input data is retrieved from NEEDS (Nikkei Economic Electronic Databank System, http ://www.nqi.co.jp/english/needs/n_top.html).

realization process within discourse unit.

3 Multilingual Generation and Voice Synthesis

The system is based on a three-stage pipelined architecture : document planning, microplanning, and surface realization. The document planning stage is independent of individual languages. The microplanning stage contains the process of conversion for lexical paraphrasing of each language. The surface realization module is dependent on each language.

3.1 Document Planning

The document planning module consists of two tasks : 'document structuring' and 'content determination'. The code fragment of each task is shown in Appendix A. In the Economy Reports' case, the output data is produced based on the previous one to three months data. Our input data is stored in Yggdrasill, the XML-DB product of Mediafusion Corporation in Japan[4], and the contents are retrieved with XPath notations and structured with DOM (Document Object Model). DOM trees are used to remove overlapping data between overview and shallow data. They are corresponding to two-stage content determination (Sripada et al., 2001). The output DTD (Document Type Definition) of this module is shown in Appendix B.

3.2 Microplanning

In the microplanning module, the XML tag and elements are replaced to produce text specification, which is based on lexical constraint in each language. More precisely, microplanning contains the following tasks : determining the detailed (sentence-internal) organization, looking at alternative ways to group information into phrases, and so on (McDonald, 2000, pp.156). This stage is implemented with SAX (Simple API for XML). Lexicalisation task according to each language and aggregation task to each sentence are modularized with SAX. The output DTD of this module is shown in Appendix C.

3.3 Surface Realization

We followed the sublanguage approach (Somers, 2000), because surface lexical expression strongly depends on each discourse structure. We implemented the surface realization stage with XSLT (eXtensible Stylesheet Language Transformations) and Xalan[5], and the output had two variations : the XHTML and VoiceXML format. The combination of *xsl* : *param* and *xsl* : *choose* command was used for lexical paraphrasing based on discourse structure constraints. In addition, context-based lexical paraphrasing is an important factor in avoiding repetitious text. We use the Java extension function in Xalan, and count the repeating element and change the expression. The completed texts of the Monthly Economic Reports are shown in Appendix D and Weather Forecasts in E.

4 Conclusions

We implemented a three-stage pipelined NLG architecture (Reiter and Dale, 2000) as XML transformation process. Our system proved the effectiveness of using XML to translate reports from a database based on the distinction between domain selection and linguistic selection.

XML is useful especially for content determination from a hierarchical structured database. If we have a time series or chronological data which is characterized by information dense at the same reference time point, our approach can be applied to NLG from such data.

Our system used the common document planner to translate into four different languages. The document planning module only depends on its input database domain. Therefore, our system makes a distinction between data selection and linguistic selection processes in order to produce reports from the DB.

References

R. I. Kittredge and A. Polguere. 2000. The generation of reports from databases. In R. Dale, H. Moisl, and H. Somers, editors, *Handbook of Natural Language Processing*, chapter 11, pages 261–304. Marcel Dekker.

D. D. McDonald. 2000. Natural language generation. In R. Dale, H. Moisl, and H. Somers, editors, *Handbook of Natural Language Processing*, chapter 7, pages 147–179. Marcel Dekker.

A. Rassinoux, R. H. Baud, C. Lovis, J. C. Wagner, and J. Scherrer. 1998. Tuning up conceptual graph representation for multilingual natural language processing in medicine. In M. Mugnier and M. Chein, editors, *Conceptual Structures : Theory, Tools and Applications*, pages 390–397. Springer LNAI 1453, Montpellier, France, 8.

E. Reiter and R. Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press.

[4]http ://www.mediafusion.co.jp/seihin/ygg/index.html

[5]http ://xml.apache.org/xalan-j/index.html

H. Somers. 2000. Machine translation. In R. Dale, H. Moisl, and H. Somers, editors, *Handbook of Natural Language Processing*, chapter 13, pages 329–346. Marcel Dekker.

S. G. Sripada, E. Reiter, J. Hunter, and J. Yu. 2001. A two-stage model for content determination. In *Proc. of the 8th European Workshop on Natural Language Generation associated to ACL 39th Ann. Meeting and 10th Conf. of the European Chapter*, pages 3–10, Toulouse, France, July 6–7.

A The Implementation Code in Document Planning Stage

A.1 Code Fragment in Document Structuring

```
if (fxoBasket.Open(Host) == 0){
  if (fxoBasket.Login(Alias, UserID, Password) == 0){
    if (fxoBasket.OpenBasket() == true){
      DomesticDemands plan1 = new DomesticDemands(xmldoc);
      Element ecoElm1 = plan1.MakePlan(year,month_i,obj,
      fxoBasket);
      PandE plan2 = new PandE(xmldoc);
      Element ecoElm2 = plan2.MakePlan(year,month_i,obj,
      fxoBasket);
      InternationalPayments plan3 = new
      InternationalPayments(xmldoc);
...
      fxoBasket.CloseBasket();
    } else {
          System.out.println("OpenBasket ----- " +
          fxoBasket.Get_Reason());
    }
    fxoBasket.Logout();
  } else {
      System.out.println("Login ----- "
      + fxoBasket.Get_Reason());
  }
  fxoBasket.Close();
} else {
  System.out.println("Open ----- "
  + fxoBasket.Get_Reason());
}
```

A.2 Code Fragment in Content Determination

```
public class DomesticDemands {
  private XmlDocument xmldoc;
  private Element subroot;

  public DomesticDemands(XmlDocument doc) {
    xmldoc = doc;
  }
  public Element MakePlan(String year, String month_i,
  String obj, JYggdrasill fxoBasket) {
    Element subroot
    = xmldoc.createElement("EconomyEvent");
    subroot.setAttribute("Type", "DomesticDemands");
    Element pc = PersonalConsumption(year,month_i,
    obj,fxoBasket);
    subroot.appendChild(pc);
    Element wg = Wages(year,month_i,obj,fxoBasket);
    subroot.appendChild(wg);
.....
  private Element PersonalConsumption(String year,
  String month_i, String obj, JYggdrasill fxoBasket) {
    Element pc
    = xmldoc.createElement("PersonalConsumption");
    int month_lll = Integer.parseInt(month_i)-3;
    int year_lll = Integer.parseInt(year);
.....
    String item1 = fxoBasket.GetDocumentFragments
    ("/EconomyData[/@Year=\"" + year_lll + "\"]
    /MonEcoRep[/@month=\"" + month_lll + "\"]
    //LivingExpenditures/text()").substring(65);
    item1 = item1.substring(0,item1.length()-17);
.....
    Element elm1 = xmldoc.createElement
    ("LivingExpenditures");
    elm1.setAttribute("Month"
    ,Integer.toString(month_lll));
    elm1.setAttribute("ComparedTo","LastYear");
    elm1.appendChild(xmldoc.createTextNode(item1));
.....
    return pc;
.....
```

B The Document Plan DTD Example

```
<?xml version="1.0" encoding="Shift_JIS" ?>
<!ELEMENT Set ( EconomyEvent+ ) >
<!ATTLIST Set Year NMTOKEN #REQUIRED >
<!ATTLIST Set Object NMTOKEN #REQUIRED >
<!ATTLIST Set Month NMTOKEN #REQUIRED >
<!ELEMENT EconomyEvent ( PersonalConsumption, Wages,
HousingConstruction, PlantandEquipmentInvestment,
MiningIndustory?, EmploymentState?, Bankruptcy?, Export?,
Import?, BalanceofPayments?,
DomesticWholesalePricesSituation?,
BargainPricesforEnterpriseSituation?,
ConsumerPricesSituation?,
FinancialSituation?, MoneySupply? ) >
<!ATTLIST EconomyEvent Type NMTOKEN #REQUIRED >
<!ELEMENT PersonalConsumption ( LivingExpenditures+,
LivingExpendituresforWorkers, LevelofConsumption,
LevelofConsumptionforWorkers ) >
<!ELEMENT LivingExpendituresforWorkers ( #PCDATA ) >
<!ATTLIST LivingExpendituresforWorkers Month
NMTOKEN #REQUIRED >
<!ATTLIST LivingExpendituresforWorkers ComparedTo
NMTOKEN #REQUIRED >
.....
```

C The Text Specification DTD Example

```
<?xml version="1.0" encoding="Shift_JIS" ?>
<!ELEMENT Set ( EconomyEvent+ ) >
<!ATTLIST Set Year NMTOKEN #REQUIRED >
<!ATTLIST Set Object NMTOKEN #REQUIRED >
<!ATTLIST Set Time NMTOKEN #REQUIRED >
<!ELEMENT EconomyEvent ( Heading, SubHeading+ ) >
<!ELEMENT Heading ( #PCDATA ) >
<!ELEMENT SubHeading ( Phrase+ ) >
<!ATTLIST SubHeading Title CDATA #REQUIRED >
<!ELEMENT Phrase ( #PCDATA ) >
<!ATTLIST Phrase Use CDATA #IMPLIED >
<!ATTLIST Phrase Class NMTOKEN #IMPLIED >
<!ATTLIST Phrase Head ( True ) #IMPLIED >
<!ATTLIST Phrase Post CDATA #IMPLIED >
<!ATTLIST Phrase Household NMTOKEN #IMPLIED >
<!ATTLIST Phrase Sbj CDATA #REQUIRED >
<!ATTLIST Phrase Rhetoric ( Sequence | Embed ) #IMPLIED >
```

```
<!ATTLIST Phrase Product CDATA #IMPLIED >
<!ATTLIST Phrase Time ( September | May | August |
October | July ) #REQUIRED >
<!ATTLIST Phrase Prep CDATA #IMPLIED >
```

D Monthly Economic Report

D.1 English Output Example

1. Domestic Demands

Personal Consumption

Living expenditures (whole) for July decreased 2.6 % compared to the same period last year, and for August a 4.1 % decrease compared to the same period last year.

When you look at the change classified by household spending, there was a 2.9 % decrease compared to the same period last year for working people in August.

The consumption level for August decreased 3.09 % compared to the same period last year.

The consumption level for working people in August decreased 2.09 % compared to the same period last year.

Wages

Income for August decreased 1.19 % compared to the same period last year for companies employing 30 or more people.

Additional allowances for August decreased 5.46 % compared to the same period last year for companies employing 30 or more people.

Real wages for August decreased 2.12 % compared to the same period last year for companies employing 30 or more people.

Housing Construction

The number of housing starts (seasonally adjusted rate) for July decreased 2.44 % compared to the last month, and a 0.53 % decrease compared to the same period last year. The number of housing starts (seasonally adjusted rate) for August decreased 0.11 % compared to the same period last year.

The floor space of new houses for August decreased 0.93 % compared to the last month, and a 2.30 % decrease compared to the same period last year.

D.2 Japanese Output Example

1. 国内需要

個人消費

個人消費は, 実質消費支出（全世帯）は前年同月比で 7 月 2.6 %減の後, 8 月は 4.1 %減となった。

世帯別の動きを見ると, 勤労者世帯では, 前年同月比で 8 月 2.9 %減となった。

消費水準指数は全世帯で前年同月比 8 月 3.09 %減 , 勤労者世帯では 同 2.09 %減となった。

賃金

賃金は, 現金給与総額（事業所規模３０人以上）は前年同月比で 8 月 1.19 %増となった。

所定外給与（事業所規模３０人以上）は前年同月比で 8 月 5.46 %増となった。

実質賃金（事業所規模３０人以上）は前年同月比で 8 月 2.12 %増となった。

住宅建設

住宅建設は, 総戸数（季節調整値）は前月比で 7 月 2.44 %減（前年同月比 0.53 %減） の後, 8 月は 0.11 %増となった。

着工床面積は前月比で 8 月 0.93 %増（前年同月比 2.30 %減） となった。

E Weather Forecast

E.1 English Output Example

The weather at three o'clock in Okinawa is cloudy throughout the archipelago .

Today's weather will be fair , but locally cloudy over the Miyako Islands . It will be clear over the Daito Islands at sunrise .

Tonight's weather will be fair throughout . It will be cloudy over the Miyako Islands , and there will be showers on the Daito Islands in the afternoon .

The outlook for tomorrow in Okinawa is fair , but locally slightly cloudy over the Daito Islands .

Tomorrow night's weather will be fair , but partly cloudy over the Yonaguni Islands .

E.2 French Output Example

Le temps à trois heures dans zone de l'Okinawa est nuageux dans toutes les parties .

D'aujourd'hui temps dans zone de l'Okinawa sera parfait , mais sera nuageux localement dans Îles de Miyako . Il sera clair dans Îles de Daito à l'aube .

Ce soir temps dans zone de l'Okinawa sera parfait dans toutes les parties . Il sera nuageux dans Îles de Miyako , et il sera pluvieux dans Îles de Daito l'après-midi .

Les perspectives pour le demain dans zone de l'Okinawa sera parfait , mais dans Îles de Daito .

Demain le temps de la nuit dans zone de l'Okinawa sera parfait , mais sera nuageux en partie dans Îles de Yonaguni .

E.3 German Output Example

Das Wetter um drei Uhr in Okinawa ist es wolkig überall in den Inseln .

Heute ist es sonnig , aber ist es vereinzelt wolkig über den Inseln Miyako . Am Sonnenaufgang wird es frei über den Inseln Daito .

Heute abend wird es sonnig überall . Am Nachmittag gibt es Duschen auf den Inseln Daito , und wolkig über den Inseln Miyako .

Morgen in Okinawa wird es sonnig , aber vereinzelt etwas wolkig über den Inseln Daito .

Morgen abend wird es sonnig , aber örtlich wolkig über den Inseln Yonaguni .