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 format¹ and synthesis function by using IBM websphere VoiceServer SDK². 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.

¹http://www.voicexml.org
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

\(^3\) In fact, our weather domain input data is the Annual Report published by the Japan Meteorological Agency (http://www.jmbsc.or.jp/online/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 dop.html).

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
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 Japan⁴, 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 struc-

---

⁴http://www.medafusion.co.jp/seihin/ygg/index.html

---

References


---

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


A The Implementation Code in Document Planning Stage

A.1 Code Fragment in Document Structuring

```java
if (fxoBasket.OpenHost() == 0) {
    if (fxoBasket.OpenBasket() == true) {
        DomesticDemands plan = new DomesticDemands(xmldoc);
        Element ecoElm = plan.MakePlan(year, month_i, obj, f xoBasket);
        PandE plan2 = new PandE(xmldoc);
        Element ecoElm2 = plan2.MakePlan(year, month_i, obj, f xoBasket);
        InternationalPayments plan3 = new InternationalPayments(xmldoc);
        ...
        f xoBasket.CloseBasket();
    } else {
        System.out.println("OpenBasket ------ + ");
        f xoBasket.OpenBasket();
    }
}
else {
    System.out.println("Login ------ ");
    f xoBasket.LogIn();
}
else {
    System.out.println("Open ------ ");
    f xoBasket.Close();
}
```

A.2 Code Fragment in Content Determination

```
public class DomesticDemands {
    private Element subroot;
    public DomesticDemands() {
        xmldoc = doc;
    }
    public Element MakePlan(String year, String month_i, String obj, Jygdrasil f xoBasket) {
        Element subroot = xmldoc.createElement("EconomyEvent");
        subroot.setAttribute("Type", "DomesticDemands");
        Element pc = PersonalConsumption(year, month_i, obj, f xoBasket);
        subroot.appendChild(pc);
        Element wg = Wages(year, month_i, obj, f xoBasket);
        subroot.appendChild(wg);
        ...
    }

    private Element PersonalConsumption(String year, String month_i, String obj, Jygdrasil f xoBasket) {
        Element pc = xmldoc.createElement("PersonalConsumption");
        int month_i11 = Integer.parseInt(month_i) - 3;
        int year_i11 = Integer.parseInt(year);
        ...
        String item = f xoBasket.getDocumentFragments("/EconomyData/@Year="+year+"/"+month_i11+"/
        +LivingExpenditure/text()", substring(65);
        item = item.substring(0, item.length()-17);
        ...
    }
```

B The Document Plan DTD Example

```
<xml version="1.0" encoding="Shift_JIS" >
  <!ATTLIST EconomyEvent Type [ EconomicEvent ] #REQUIRED >
  <!ATTLIST EconomyEvent Type [ Household, Financial, SubEconomy, Economic ] #IMPLIED >
`
```

C The Text Specification DTD Example

```xml
<xml version="1.0" encoding="Shift_JIS" >
  <!ELEMENT EconomyEvent ( Heading, SubHeading* ) >
  <!ATTLIST EconomyEvent Heading [ "#PCDATA" ] #REQUIRED >
  <!ATTLIST EconomyEvent SubHeading [ Phrase ] #IMPLIED >
  <!ATTLIST EconomyEvent Title [ CDATA ] #REQUIRED >
  <!ATTLIST EconomyEvent Phrase [ CDATA ] #IMPLIED >
  <!ATTLIST EconomyEvent Phrase [ Phrase ] #IMPLIED >
  <!ATTLIST EconomyEvent Household [ CDATA ] #REQUIRED >
  <!ATTLIST EconomyEvent Phrase Rhetoric ( Sequence | Embed ) #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，労働者世帯では，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，労働者世帯では 同 2.09 %減となった。

貯金は，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。
実質貯金（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.


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.