Automatic feeding of translation memory tools

Inés Jacob
Facultad de Ingeniería
ines@eside.deusto.es
fax: 34(9) 44 139101

Joseba Abaitua, Josu Gómez
Facultad de Filosofía y Letras
abaitua@fil.deusto.es
fax: 34(9) 44 139087

Universidad de Deusto
Avda. de las Universidades, 24
48007 Bilbao

Abstract

We present a method for the automatic extraction of aligned segments as a way to speed-up the process of translation-memory development. Parallel documents are captured from bilingual sites in the Web and processed in two steps. First, relevant information is extracted from downloaded files and transformed into a simplified TEI format. In a second step, parallel TEI files are converted into single files in TMX, which contain large collections of translation segments in a format that allows for the automatic feeding of translation memory tools.

1. Introduction

The development of methods and tools that accelerate the translation of multilingual digital contents is a matter of great importance in the information society. Normally translators work with documents that are related in one way or another with other documents that they have previously translated. Translators often recognize fragments that they would like to consult or even reuse, but which they fail to locate in the different memory devices of their computers.

Translation Memory (TM) systems are tools that help translators carrying out this task in a manner that is respectful with the way they normally work. These systems make it possible to store and retrieve bilingual fragments, which normally correspond to previously translated paragraphs or sentences. TMX (Translation Memory eXchange format) is the standard method that most system use to exchange translation data among tools.

The DELi research group at the University of Deusto specializes in the mechanization of the different stages that are involved in the process of generating translation memories in TMX format (Abaitua et al. 2002. Jacob et al. 2002).

Recent projects cover the following issues:

- analysis, selection, and integration of different software-tools that support import and export of multilingual contents in TMX and TBX;
- work-flow management tools for multilingual contents based on TMX and XLIFF; and
- multilingual web update and maintenance.

As a result, procedures for the exchange of translation and localization resources have been defined with application in content-management for multilingual web sites.

2. File-capture from the Web

As a way to benchmark-test these techniques, were extracted from the Web documents belonging to several years of the official bulletins of our local institutions in the Basque Country. Texts from these bulletins contain a particular type of language, with a predominance of legal and administrative terminology, a distinctive class of registers, and a highly structured distribution of contents. The documents published in Internet normally contain monolingual texts, with the versions in Spanish and Basque in different separated files.

In some cases, the original file format was plain text, with no tags, in others we were able to get files in HTML. (There were also big collections of files in PDF format, which we did not process.) This diversity of formats in origin is unavoidable, and requires the development of particular program modules for each particular source.

3. Information Extraction and TEI tagging

TEI markup has been adopted because it is a standard format for electronic text interchange, but also because it permits a precise cataloguing method. TEI-headers provide an effective way to manage the corpus. Furthermore, since TEI markup can be expressed in XML, the corpus can be browsed by means of common web-exploring facilities. Our particular corpus has been described through a simplified tag-set (TEI-Bi) inspired in TEI-Lite.

Each document has been saved separately and converted into a XML file with the basic structure shown in Figure 1.

```
<teiHeader>...</teiHeader>
<text>
 <front> [Information from the front-page, up to the last title]
 </front>
 <body> [The text properly, including place and date] </body>
 <back> [place, date, signature, and control number] </back>
</text>
```

Figure 1. Document structure

The header (<teiHeader>) assigned to each documents contains the elements of Figure 2.

```
<teiHeader>
    <fileDesc>
        <titleStmt>
            <title>
            <author>
            <principal>
        <publicationStmt>
            <publisher>
        <sourceDesc>
            <bibl>
    <encodingDesc>
        <projectDesc>
        <classDecl>
            <taxonomy>
                <bibl>
    <profileDesc>
        <creation>
        <langUsage>
        <textClass>
            <classCode>
```

Figure 2. Structure of the document header

In order to meet the requirements of the DTD defined in TEI-Bi, it is necessary to extract from the texts a series of relevant data, such as the values of tags as title, author, editor, language, etc. The uniform structure of downloaded documents facilitates this task. In Figure 3 we show some of the automatically identified elements which will feed the teiHeader.

N	*Elemento de información*	*Ejemplo*
1	Título del Boletín	Boletín Oficial de Gipuzkoa
2	Número	Número 23
3	Fecha	Fecha 01-02-2001
4	Página	Página 1843
5	--	--
6	Num y Administración	7 Administración Municipal
7	Nombre Departamento	Ayuntamiento de Errenteria
8	Descripción de la Orden	Aprobación e información pública...
9	--	--
10	Departamento Superior (puede omitirse o ser doble)	Ayuntamiento de Errenteria
11	Departamento Inferior (puede omitirse)	Rentas y Exacciones
12	Tipo o Título de la Orden	Anuncio
13	Texto	TEXTO
14	Lugar	Errenteria,
15	Fecha	a 19 de enero de 2001
16	Firma o Firmas	El Delegado del Area de Hacienda...
17	Control	(392)(963)
	Tabla (existe a veces)	

Figure 3. Automatically identified elements

The text is marked-up in accordance with the following template. The numbers in parentheses identify elements as shown in Figure 3.

```
<front>
(1) <head type="main">
           (2) <head type="num">
(3) <docDate type="main">
(4) <head type="page">
     (6) <head type="Admin">
           (7) <head type="Dep">
           (8) <head type="law">
           (10) <head type="DepSup">
           (11) <head type="DepInf">
(12) <head type="abbreviated">
</front>
<body>
           (13)  <p>
                 <p><table></p>
                 <list type="ordered">, <list type="bulleted">
</body>
<back>
     (14) <head type="place">
           (15) <docDate type="sub">
           (16) <docAuthor>
(17) <head type="control">
</back>
```

For each document, the elements <body> (<p> and <list>) are ordered by means of an id attribute whose value will be 1, 2, etc.. The element <table> will always appear inside a <p>, and will be assigned no number.

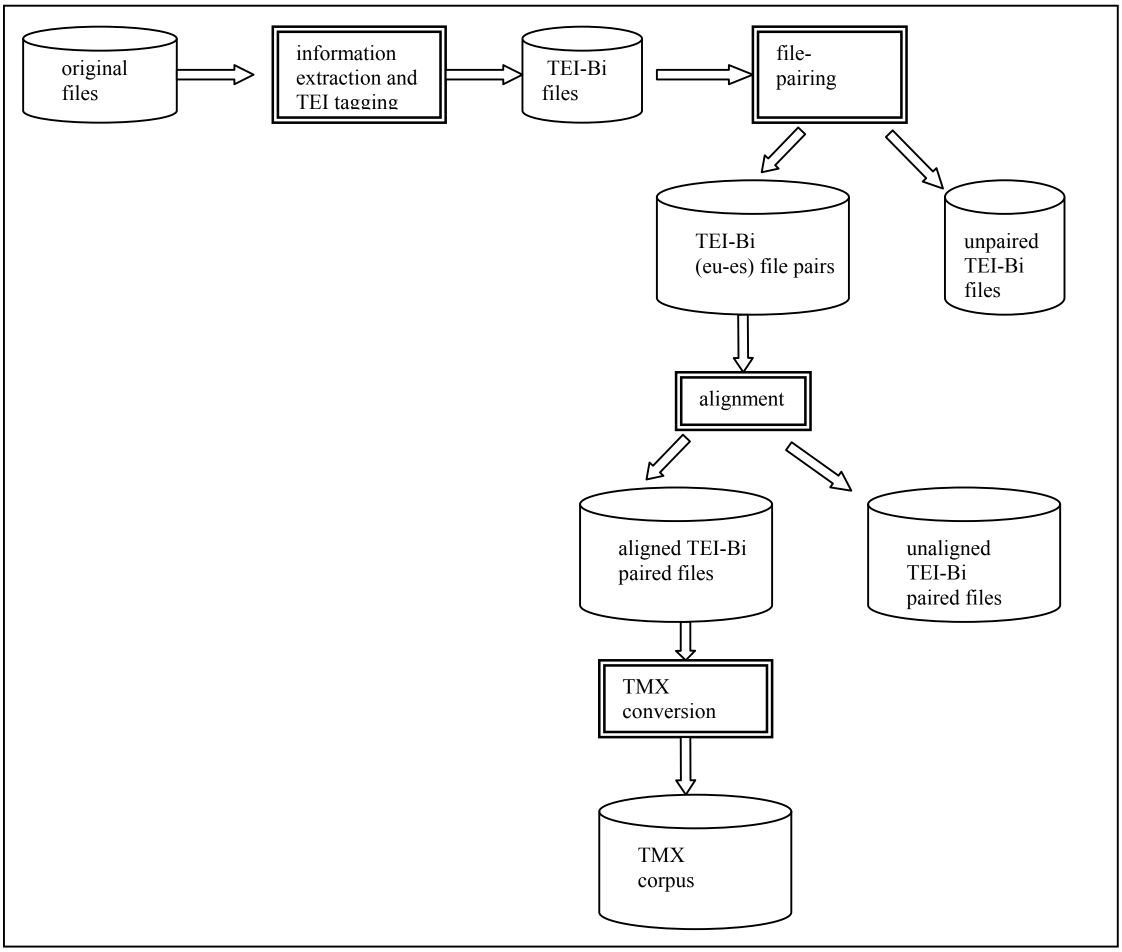
4. Alignment and TMX markup

Once documents are assigned a header and a basic set of TEI tags, original files and their translations are paired. These paired files are automatically processed so as to recognize and annotate every parallel segment. The output of this process are two files with aligned segments explicitly marked-up. Alignment takes place at the level of well-defined textual units, mainly paragraphs and headings. If for two paired files the number of segments does not match, and alarm is activated and the output set aside for manual revision.

The final format chosen for the aligned corpus is TMX because it facilitates a quick reutilization of the material. Once alignment of TEI documents has been carried out, conversion into TMX is carried out through a series of filters.

TMX files also contain a header and a body. The header provides information regarding the file's references (author, title, language, dates, publisher, etc.). The body is made up of equivalent fragments that are called units of translation, marked-up as <UT>. These translation units correspond to the segments already aligned in the TEI files.

The architecture of the system is as follows:


original files
information extraction and TEI tagging
TEI-Bi files
file-pairing
TEI-Bi (eu-es) file pairs
unpaired TEI-Bi files
alignment
aligned TEI-Bi paired files
unaligned TEI-Bi paired files
TMX conversion
TMX corpus

5. Results

The output of the process is an automatically obtained translation memory in TMX. It is important to take into account that each new source of bilingual information will have specific characteristics with regard to the structure of the documents, to the storage-format, and to the filing structure. For this reason, the tools that render TEI-Bi formats will have to adapt to each particular information-source. However, from that stage onwards, the acquisition TMX memories can be carried out with any further adaptation.

6. XLIFF for multilingual content managment

The definition of the XML Localisation Interchange File Format (XLIFF) started in 2000 by a group of localization suppliers and tools vendors, including: Oracle, Novell, IBM/Lotus, Sun MicroSystems. This format is based on the principles defined as OpenTag, and adopts many of the ideas later developed for TMX, with innovations that facilitate the exchange of the information that must be translated. The first draft of XLIFF, version 1.0, was published in May 2001. The development of the format was moved under the aegis of OASIS in December 2001, and the XLIFF version 1.0 became an OASIS Committee Specification at the beginning of 2002. The version 1.1 is being worked on currently.

We are currently studying the adoption of XLIFF as a way to facilitate the work-flow management of multilingual contents in the different stages of the translation and location processes.

7. References

Joseba Abaitua, JosuKa Díaz, Josu Gómez, Inés Jacob, Garikoitz Araolaza, Luistxo Fernández. 2002. SARE-Bi: Gestor de documentación multilingüe sobre XML/TEI. *Procesamiento de Lenguaje Natural* 29: 313-314.

Inés Jacob, Josu Gómez, Joseba Abaitua. 2002. Obtención de corpus bilingües para la alimentación automática de gestores de memorias de traducción. II Congreso Internacional de Traducción Especializada: La traducción científica. Universitat Pompeu Fabra, Barcelona

TEI: http://www.tei-c.org

TMX: http://www.lisa.org/tmx

XLIFF: http://www.opentag.com/xliff.htm