

Glossary Embedding System for English & Japanese

Pingkui Hou, Naoyuki Tokuda, R&D Center, SunFlare Co.Ltd, Japan

In the translation field, glossary can be used for words and phrases that have a specific meaning within the corresponding domain knowledge. Keeping the glossary identical is very important for the localization and internationalization in order to deliver the prorogating information. Traditionally, these works always was done by the experienced translator manually, which caused the inefficiency. STPP(SunFlare Term PreProcessor) is another way to accomplish these work with a higher efficiency.

Given a document, STPP can extract the important glossaries –nouns and compounds, and pre-embed them into the document after they were confirmed by the experienced translator.

The characteristics of STPP include:

- (1) XML format: Information prorogated internally is based on XML. XML is becoming more and more popular both as the web application and the office application because of its compatibility and easy delivery, etc.
- (2) Multi file type supported: PDF, MS Office, LaTeX are the three most important file format of the delivered document. With the OpenOffice's development, more and more people turned to use the OO. STPP can import these most popular file formats by using the file filters – which were the components of StarSuites of SunMicro at present time.
- (3) Multi language: Both English and Japanese documents can be processed by using the corresponding natural language parser.
- (4) Web-application integrity: With the development of Internet, any application can grow up into a GAINT with its Internet accessibility. With the Web-application integrity, documents can be processed remotely under the certain security protocols.
- (5) Flexibility: STPP can be expanded to support other language documents easily.

STPP can be described in terms of five basic modules:

- (1) File filters : which identify the file type and convert it into XML-based format.
- (2) Natural Language Parser: English documents and Japanese documents will be processed by Brill's parser and CHASEN respectively.
- (3) Glossary Extraction: STPP concerns with nouns and compounds by using the Bi-gram algorithm [N. Mori].
- (4) Embedding Module: which embeds the translated glossary into the document.
- (5) Glossary Database: including the indexing agent and updating manager.

