A Hybrid Approach to Multilingual Text Processing: Information Extraction and Machine Translation

Chinatsu Aone, Hatte Blejer
Systems Research and Applications Corporation (SRA)
2000 15th Street North
Arlington, VA 22201
aonec@sra.com, blejerh@sra.com

Mary Ellen Okurowski, Carol Van Ess-Dykema
Department of Defense
9800 Savage Rd.
Ft. Meade, MD 20755-6000
meokuro@afterlife.ncsc.mil, cjvanes@afterlife.ncsc.mil

Abstract

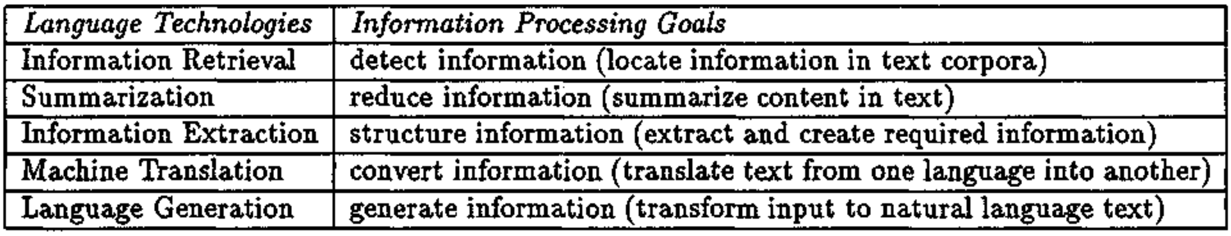
In this paper, we discuss the effectiveness of a particular selection and sequencing of two emerging language technologies, namely information extraction (IE) followed by machine translation (MT), and report and evaluate our experiment results. The experiments have shown that different template slot types affect the quality of MT results, and that MT quality of some slot values is actually helped by IE while others, though less frequent, suffer from a lack of contextual information as the result of IE.

1 Introduction

As language technologies emerge, we must assess their applicability to various tasks and determine how they can be used in a complementary fashion. In this paper, we will focus on the joint use of information extraction and machine translation. We will present empirical data which demonstrates the effectiveness of this combination of technologies for an example task frequently performed in government and commercial sectors.

2 Language Technologies

In order to assess the applicability of emerging language technologies, we propose the following functional schemata. Our classification is based on the fact that different types of information processing are required for different goals. These technologies include information retrieval (IR), summarization (SUM), information extraction (IE), machine translation (MT), and language generation (GEN). Table 1 shows these technologies and their respective goals.



Language Technologies	*Information Processing Goals*
Information Retrieval	detect information (locate information in text corpora)
Summarization	reduce information (summarize content in text)
Information Extraction	structure information (extract and create required information)
Machine Translation	convert information (translate text from one language into another)
Language Generation	generate information (transform input to natural language text)

Table 1: Language Technologies and their Goals

3 From Multilingual Information Extraction to Machine Translation

Various tasks in both government and commercial sectors require applications of different types of language technologies. For one task, the user might want to first detect foreign language texts of interest, and then fully translate selected texts into English. For another task, the user may be interested in reviewing summarized versions of sets of texts. Here we describe a different task. The task is characterized by the following parameters:

- processing a large number of foreign language on-line texts
- extracting predefined generic types of information
- structuring extracted information for downstream applications (e.g. visualization, database query)
- providing information to users, some of whom do not know the foreign language

These parameters guide us to select and sequence two language technologies in our hybrid approach. We propose applying information extraction, specifically *multilingual* information extraction (MIE), followed by machine translation (cf. [3]). That is, we apply information extraction to texts in foreign languages and then apply machine translation to the output of extraction. Figure 1 shows two possible configurations of the two technologies.

The MIE => MT configuration exploits the maturities of the two technologies in the most economical and optimal way. It is more economical because, assuming that Japanese and English MIE systems are similar in quality[1], the time to do the present task (i.e. creating templates in I English from texts in foreign languages) depends on the time to perform MT. Then, in the case of I MIE => MT, an MT system has to translate only the values of templates while in the case of MT => MIE, an MT system has to translate the entire texts.[2] The proposed configuration is also more optimal. Because an MIE system analyzes the text and structures the information, consequently 1 it constrains and simplifies the input to an MT system, improving the quality of MT output. I

In the following, we briefly describe MIE and MT systems, and then discuss the proposed hybrid 1 approach. I

I

[1] We also assume that there are humans in the loop who correct the output of MIE and MT systems in either configuration.

[2] In addition, editing of the MT template output is clearly easier than that of the entire text.

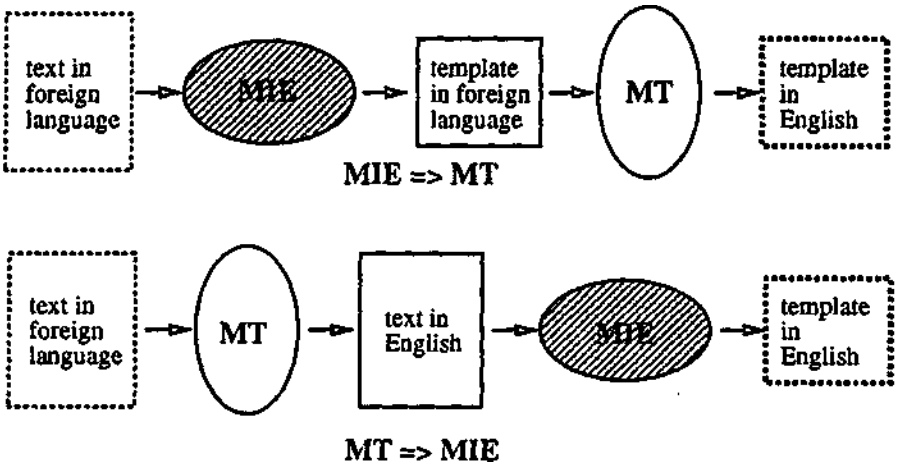


Figure 1: Two Configurations of MIE and MT

3.1 Multilingual Information Extraction

An information extraction system (cf. [1]) automatically analyzes texts as directed by the user's requirements. In particular, an IE system locates and extracts relevant information from texts in foreign languages, and structures the extracted information in user-defined templates in order to feed such downstream applications as visualization tools, database queries, and automated trend analysis tools. For example, an IE system extracts information about specific events and their associated properties such as organizations, people, time, locations, etc. In addition, an IE system can infer and add additional information. For most downstream applications, the users verify the output of an IE system.

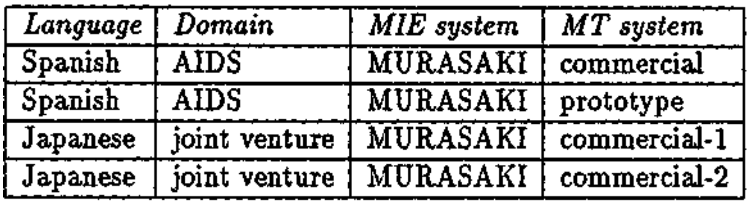
A *multilingual* information extraction system extracts from foreign language texts. While MIE systems access language-specific knowledge sources in processing foreign language texts, some of the MIE systems represent the extracted information internally in a language-neutral way using an interlingua. The MURASAKI system (cf. Aone *et al.* [2, 4]) is an example of such a system. Other MIE systems can only output extracted information in the given language of the text.

A portable and extensible MIE system consists of a core language processing engine and various knowledge sources particular to domains and/or languages (cf. Hobbs [6]). The former typically consists of preprocessing, syntactic, semantic, and discourse modules, and the latter includes lexicons, grammars, and knowledge bases.

3.2 Machine Translation

A machine translation (MT) system automatically converts foreign language texts in a source language into a target language. There are currently a number of approaches to machine translation. For example, some MT systems primarily use statistical information derived from aligned corpora to perform translation (cf. Brown *et al.* [5]), while others use primarily linguistic information (cf. Knight [7]). Still others use a combination of the two approaches (cf. Yamron *et al.* [8]).

The users often pre-edit the input text, and also post-edit the output of the system. In addition, some systems allow machine-aided translation, i.e. the user helps the system translate the text by



Language	*Domain*	*MIE system*	*MT system*
Spanish	AIDS	MURASAKI	commercial
Spanish	AIDS	MURASAKI	prototype
Japanese	joint venture	MURASAKI	commercial-1
Japanese	joint venture	MURASAKI	commercial-2

Table 2: Four Test Settings

responding to system queries.

3.3 A Hybrid Approach

In order to extract user-defined information from a source text and output it as structured information in a target language, we propose combining SRA's multilingual information extraction system MURASAKI with (1) commercially available MT systems and (2) a government-supported MT prototype system.

We have previously applied the MURASAKI system to MIE tasks in multiple domains and languages. To demonstrate our hybrid approach, we have chosen two language-domain pairs, i.e. Spanish/AIDS and Japanese/joint-venture.

In the AIDS task, the system analyzes Spanish newspaper articles which report AIDS disease statistics, and extracts information such as the number of AIDS and HIV incidences reported in particular countries during particular time frames. For the joint venture task, MURASAKI analyzes Japanese newspaper texts which report joint ventures between organizations, and derives information such as the organizations and people participating in the joint venture and purposes of the newly formed tie-up relationships. The system processes the foreign language input texts *directly,* and outputs structured information in that foreign language (e.g. Spanish, Japanese). There is no "machine translation" involved in this process.

The structured data we create here is a template which consists of slots and their values. The users define the slots, and the system fills the slot values. These slot values can be normalized fills (e.g. dates), set fills (i.e. a finite list of possible values), and string fills (i.e. values derived from the original text). The sample Japanese text and template are shown in the appendix.

4 Experiments and Results

In order to assess the effectiveness of the hybrid approach, we have designed tests which take MURASAKI output and apply MT systems to translate the Spanish and Japanese slot values into English. We have taken two MT systems for each language and had them process 20 templates for each domain, totaling 80 tested templates. Table 2 shows four different tests settings.

The quality of MT varied greatly according to the types of template slot values, but not according to different MT systems. Translation of *set fills* (e.g. "país," "ciudad"), which is a large portion of input to MT, was generally accurate. Because set fills are finite lists of choices and usually one word, they are trivial to translate. However, there were a few cases where contextual information, as exists in a whole text, would have helped avoid mistranslation. For example, "diario" for the type of organization slot was translated into "diary" although it meant "newspaper" in the text. Given the trivial nature of translating set fills and the occasional need for contextual information, we believe that letting an MIE system output set fills in English directly through a simple set fill conversion table lookup, which MURASAKI can do currently, is a more realistic option.

Normalized fills, which were normalized dates (e.g. 26 AGO 88, 881201) in our experiments, were not translated correctly by MT systems because their normalization formats were specific to the tasks. These fills should be also handled directly by an MIE system, since it can analyze dates and output them in any format.

Overall, the MT quality of *string fills* were worse than those of set fills. This was especially the case with longer phrases and sentences in both Japanese and Spanish and Japanese proper names. When names are unknown in Japanese, the MT systems usually left them in Japanese in the English translation.[3]

Analysis of string fills revealed that the effect of contexual information on MT differed for two types of string fills. First, string fills which are proper nouns, such as names of people, organizations, and locations, sometimes needed *local context* to identify that these phrases are actually proper nouns, and not other parts of speech. For example, Spanish personal names "Alceni Guerra" and "Mayté Paredes" were translated into "Alceni War" and "Mayte Walls", where the last names were mistaken for common nouns. When provided the whole texts, however, the same MT system recognized the phrases as proper names and translated them correctly. In Japanese, a company name "Kanematsu" was incorrectly translated into "And pine" and a location name "Marirando (i.e. Maryland)" into "Husband land" because the MT system lacked contextual information. Such local context is, however, actually present in the slot names of the templates, such as PERSON-NAME, ENTITY-NAME, and LOCATION. We think that it would be ideal if an MT system can take semantic types of phrases (e.g. person, organization, location) as an optional parameter and provide for top-down expectation to help translate the phrases appropriately.

The lack of contexual information did not affect the MT quality of the other type of string fills, namely phrasal string fills. In fact, MIE helped improve MT quality by delimiting input and thus reducing the linguistic complexity. Template fill rules of given MIE tasks specify allowable linguistic complexity for phrasal string fills. For example, sentence 1 below is the result of MT after MIE, and sentence 2 is the result of MT before MIE. The only difference in the Japanese input sentences was that the former did not contain an adverb "shourai-wa" because it was not necessary as MIE output.[4] It is easy to see that a small simplification of an input sentence made a drastic improvement in the MT quality.

1. The railroad ticket and amusement park of a private-line system, the entrance ticket of sight-seeing institution, etc. are added to a menu.

2. It is a private-line system in future. A railroad and amusement park, the entrance ticket of sight-seeing institution, etc. are added to a menu.

[3] It seems that the MT systems we used need a capability to transliterate Japanese proper names automatically into English.

[4] See the appendix for the template and text which the input sentences for 1 and 2 come from.

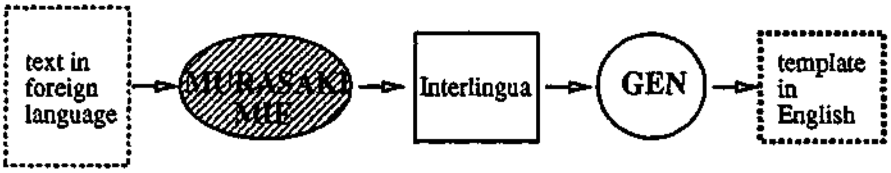


Figure 2: MIE => GEN

5 Current and Future Work

In order to assess the sequencing of the two technologies, we are currently conducting and evaluating experiments where we reverse the sequence of the technologies in the subset of the data. We take Japanese joint venture texts and translate them using a commercial MT system. We then take the English translation and extract and structure the data using the MURASAKI system. We will be able to compare the results of this test with the data processed with the hybrid approach. Given the necessity to have humans in the loop, we expect that editing the MT output of whole texts takes more time than editing that of templates, thus strengthing our claim for the economic advantage of the MIE => MT sequence.

We will also evaluate a combination of an MURASAKI MIE system, which can create internal interlingua representations, with a language generation prototype system as an alternative to MT systems (cf. Figure 2). A language generation system will accept as input the interlingua representation from an MIE system, and output phrases or sentences in a target language. The advantage of using a generation system rather than an MT system is that the results of analyzing the whole text for information extraction purposes are already available in an interlingua representation as input to the generation system.

References

[1] Advanced Research Projects Agency. *Proceedings of Fifth Message Understanding Conference (MUC-5).* Morgan Kaufmann Publishers, 1993.

[2] Chinatsu Aone, Hatte Blejer, Sharon Flank, Douglas McKee, and Sandy Shinn. The Murasaki Project: Multilingual Natural Language Understanding. In *Proceedings of the ARPA Human Language Technology Workshop,* 1993.

[3] Chinatsu Aone, Hatte Blejer, Sharon Flank, Mary Ellen Okurowski, and Carol Van Ess-Dykema. MURASAKI: Multilingual Extraction and Targeted Translation. Presented at *Symposium on Advanced Information Processing and Analysis,* 1994.

[4] Chinatsu Aone, Sharon Flank, Paul Krause, and Doug McKee. SRA: Description of the SOLOMON System as Used for MUC-5. In *Proceedings of Fourth Message Understanding Conference (MUC-5),* 1993.

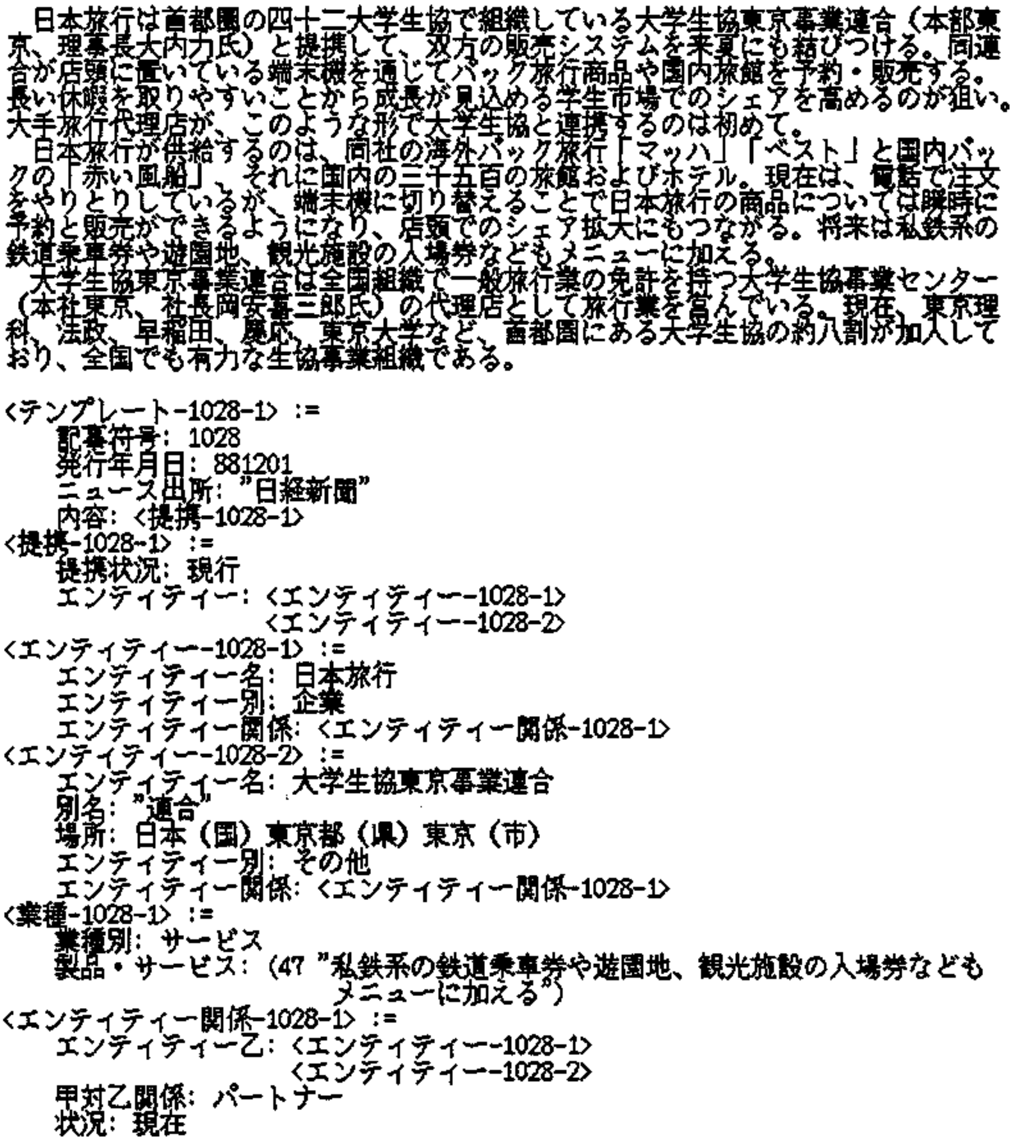
[5] Peter Brown, Stephen Delia Pietra, Vincent Delia Pietra, and Robert Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics,* 19(2), 1993.

[6] Jerry Hobbs. The Generic Information Extraction System. In *Proceedings of Fifth Message Understanding Conference (MUC-5),* 1993.

[7] Kevin Knight. Building a Large Ontology for Machine Translation. In *Proceedings of the ARPA Human Language Technology Workshop,* 1993.

[8] Jonathan Yamron, James Cant, Anne Demedts, Taiko Dietzel, and Yoshiko Ito. The Automatic Component of the LINGSTAT Machine-Aided Translation System. In *Proceedings of the ARPA Human Language Technology Workshop,* 1994.

APPENDIX



　日本旅行は首都圏の四十二大学生協で組織している大学生協東京事業連合（本部東京、理事長大内力氏）と提携して、双方の販売システムを来夏にも結びつける。同連合が店頭に置いている端末機を通じてパック旅行商品や国内旅館を予約・販売する。長い休暇を取りやすいことから成長が見込める学生市場でのシェアを高めるのが狙い。大手旅行代理店が、このような形で大学生協と連携するのは初めて。

　日本旅行が供給するのは、同社の海外パック旅行「マッハ」「ベスト」と国内パックの「赤い風船」、それに国内の三千五百の旅館およびホテル。現在は、電話で注文をやりとりしているが、端末機に切り替えることで日本旅行の商品については瞬時に予約と販売ができるようになり、店頭でのシェア拡大にもつながる。将来は私鉄系の鉄道乗車券や遊園地、観光施設の入場券などもメニューに加える。

　大学生協東京事業連合は全国組織で一般旅行業の免許を持つ大学生協事業センター（本社東京、社長岡安喜三郎氏）の代理店として旅行業を営んでいる。現在、東京理科、法政、早稲田、慶応、東京大学など、首都圏にある大学生協の約八割が加入しており、全国でも有力な生協事業組織である。

```
<テンプレート-1028-1> :=
    記事符号: 1028
    発行年月日: 881201
    ニュース出所: "日経新聞"
    内容: <提携-1028-1>
<提携-1028-1> :=
    提携状況: 現行
    エンティティー: <エンティティー-1028-1>
                    <エンティティー-1028-2>
<エンティティー-1028-1> :=
    エンティティー名: 日本旅行
    エンティティー別: 企業
    エンティティー関係: <エンティティー関係-1028-1>
<エンティティー-1028-2> :=
    エンティティー名: 大学生協東京事業連合
    別名: "連合"
    場所: 日本（国）東京都（県）東京（市）
    エンティティー別: その他
    エンティティー関係: <エンティティー関係-1028-1>
<業種-1028-1> :=
    業種別: サービス
    製品・サービス: (47 "私鉄系の鉄道乗車券や遊園地、観光施設の入場券なども
                     メニューに加える")
<エンティティー関係-1028-1> :=
    エンティティー乙: <エンティティー-1028-1>
                      <エンティティー-1028-2>
    甲対乙関係: パートナー
    状況: 現在
```