Information Extraction across Linguistic Barriers*

Megumi Kameyama
Artificial Intelligence Center
SRI International
333 Ravenswood Ave., Menlo Park, CA 94025, U.S.A.
megumi@ai.sri.com

Abstract

Information extraction (IE) systems have been tailored to extract fixed target information from documents in a fixed language. In order to be truly useful for information analysts, the target information must be user-definable and the source documents should cover multiple languages. We will map out the path toward such open-target multilingual IE systems, identifying necessary technological breakthroughs along the path. We also discuss a Japanese-English named entity extraction system under development, which represents a case of the next step along the path.

Introduction: Toward Multilingual Information Extraction Systems

The natural language processing field has witnessed a rapid development of the information extraction (IE) technology since the early 90's, driven by the series of Message Understanding Conferences (MUC's) in the government-sponsored TIPSTER program.[1] This technology enables a rapid, robust, and automatic extraction of certain predefined target information from real-world on-line texts or speech transcripts accessible through computer networks.

Information analysts, whose task is to keep track of changing states of affairs about particular topics such as microelectronic products and international terrorist activities, can use the IE technology for accomplishing their tasks more efficiently and effectively.

IE systems, however, have so far been tailored to extract fixed target information from documents in a fixed language. In order for the IE technology to be truly useful for information analysts, the target information must be user-definable, or 'open,' and it should also obtain information from documents in multiple languages.

In this paper, we will map out the path toward such open-target multilingual IE systems, identifying necessary technological breakthroughs along the path. We also discuss a Japanese-English named entity extraction system under development, which represents an IE system that lies in the immediate future along the path.

Information Extraction Technology

Given a set of source documents, the input to an IE system is a description of the target information type, and the output is a set of target information instances found in the source documents.

The target information, typically of the form "who did what to whom where when," is extracted from natural language sentences or formatted tables, and fills parts of predefined template data structures with slot values. Partially filled template data objects about the same entity or event instances are then merged to create a network of related data objects. These template data objects depicting instances of the target information are the raw output of information extraction, ready for a wide range of applications such as database updating or summary generation. An IE output can also take the form of SGML or other types of markups on the source documents, which need not go through template data objects.

IE systems, however, have so far been tailored to extract fixed target information from documents in a fixed language, through a short but intense customization work by IE system experts. We can characterize these first-generation IE systems as systems for *monolingual* information extraction with *closed* target information. See Figure 1.

Analysis of MUC evaluation results has led to a clearer understanding of the strengths and weaknesses of the technology. The MUC-6 evaluation results (Sundheim 1995) showed that name recognition is largely a solved problem, with a human-like performance in the same task (higher F-measure around 95%). This IE subtask is thus ready for real-world applications. Extracting template entities (higher F-measure around 75%) and recognition of

*To appear in the *Working Notes for the Workshop on Cross-Language Text and Speech Retrieval*, AAAI Spring Symposium Series, Stanford, CA, 1997.

[1] The TIPSTER web page is at http://www.tipster.org

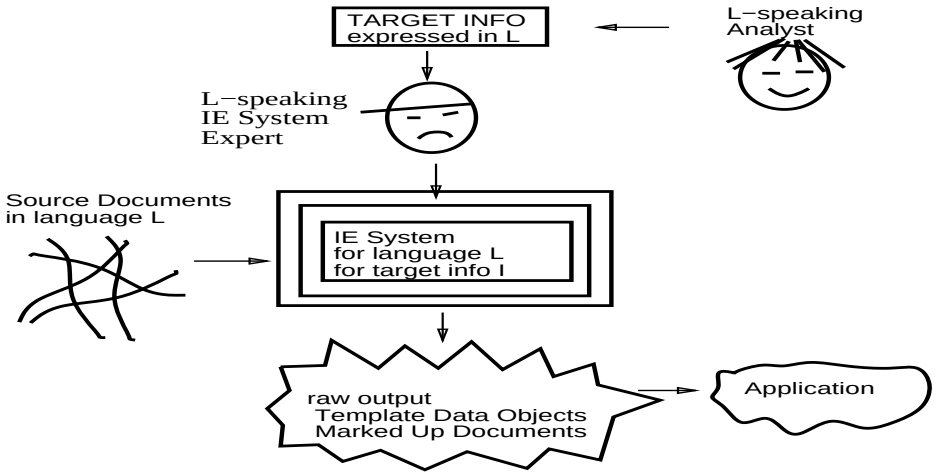

Figure 1: Monolingual Information Extraction with Closed Target Information

coreference links among noun phrases (higher recall around 60%, precision around 70%) are moderate challenges. Recognition of domain-dependent scenarios is the biggest challenge (higher F-measure around 55–60%). The level of difficulty in IE tasks reflects the amount of semantic and discourse-pragmatic relationships among partially described objects and events that the system must figure out.

I will take a long-term perspective here, assuming that eventual IE systems need to do well in domain-dependent scenario extraction. Name recognition, template entity extraction, and coreference recognition are largely domain-independent tasks, each of which can be used for real-world applications, but they are primarily important as necessary component tasks within scenario extraction.

Multiple sites working on IE have converged on the use of finite-state linguistic patterns to identify relevant descriptions of the target information. This approach has proven effective for real-world written texts such as newspaper articles. We also found that this approach is effective for errorful transcripts of spoken input. SRI's FASTUS has been used for an automatic summarization of human-human spontaneous spoken dialogues about conference room scheduling, and achieved a recall of 77.7% and precision of 82.5% on a blind test set of 75 dialogues (Kameyama & Arima 1993; 1994; Kameyama, Kawai, & Arima 1996; Kameyama 1995).

Open-Target Information Extraction

A closed-target IE system is not very useful because analysts would want to define new target information or modify parts of the existing target information as their needs change over time. An IE system, therefore, must be customizable with open target information within a sufficiently broad domain. These second-generation IE systems will do monolingual information extraction with *open* target information. See Figure 2.

An open-target IE system crucially needs a component that turns the analyst's target information expressed in plain English into a wide range of patterns that capture the various ways in which this information may be described. For instance, the analyst can define the target information as "Companies form tie-up relationships with companies," and this basic sentence triggers a generation of patterns for its lexical and morphosyntactic variants. Some examples follow.

Base <CO> "form" "tie-up relationship" "with" <CO>, e.g., *Meiji Nyugyo formed a tie-up relationship with Colonial Beef*

Nominal1 "tie-up relationship" "between" <CO> "and" <CO>, e.g., *the tie-up relationship between Meiji and Colonial ended in May*

Nominal2 <CO> "'s" "tie-up relationship" "with" <CO>, e.g., *Meiji's business tie-up relationship with Colonial ...*

These syntactic variants include Passivization, Relativization, Nominalization, Clefting, and Infinitive and Gerundive forms of the basic sentence pattern. Such syntactic transformation components have been added to the recent IE systems to facilitate the process of defining new target information (Appelt *et al.* 1995; Grishman 1995; Hobbs *et al.* 1996a). An open-target IE system also needs to know synonyms and paraphrases. For instance, "signing a tie-up contract" must be one of the basic expressions for "forming a tie-up relationship." In order to form correct surface variants of particular verbs, the system also needs to know their case frames (Nominative, Accusative, etc.) with corresponding grammatical functions (Subject, Object, etc.) and thematic roles (Agent, Patient, Experiencer, etc.).

In short, an open-target monolingual IE system must embody the monolingual knowledge of an IE system

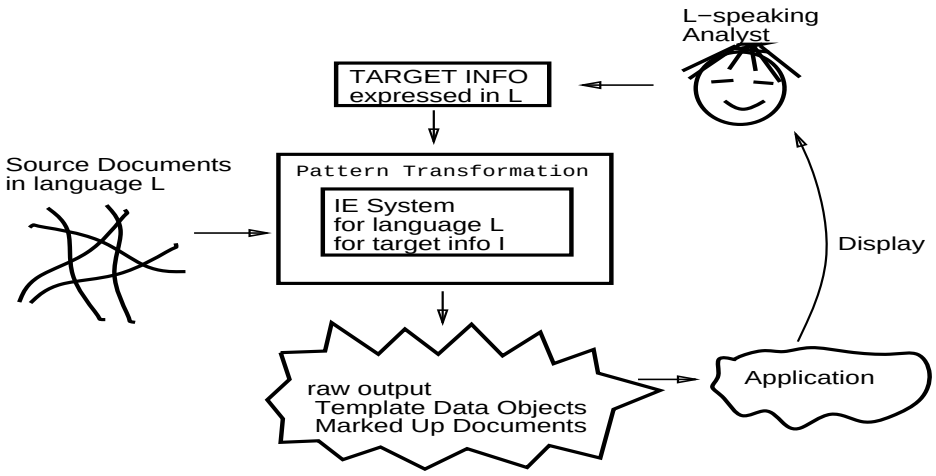

Figure 2: Monolingual Information Extraction with Open Target Information

expert, who can define specific patterns for all the lexical and morphosyntactic variants that could potentially describe examples of a given target information type.

For a multilingual IE system discussed next, this open-target capability of monolingual systems becomes crucial because IE system experts for a large number of languages may not be readily available every time the target information changes.

Multilingual Information Extraction

The performance of a number of Japanese MUC-5 systems in 1993, including SRI's Japanese Joint Venture FASTUS (Appelt *et al.* 1993a), demonstrated that the basic IE technology initially developed for English sources is portable to a language very different from English with comparable results. The government-organized Multilingual Entity Task (MET) in 1996, in which SRI participated with the Japanese FASTUS (Kameyama 1996), also demonstrated that name recognition in foreign language texts achieves a comparable high performance level (higher F-measure exceeding 90%). These demonstrations, however, stopped short of breaking linguistic barriers. It did not enable, for instance, an English-speaking analyst to access information obtained from non-English language sources. For this to happen, we need to combine two NLP technologies — information extraction and machine translation (MT).

There are several possible combinations of IE and MT. One possibility is to first translate the entire source documents into the analyst's language L_a before doing L_a-monolingual IE. This MT-IE combination is likely to suffer from all the drawbacks of full-scale MT — it is difficult, time-consuming, and costly. Full-scale discourse translation requires complex inferences (Hobbs & Kameyama 1990) and context-dependent resolutions of mismatches (Kameyama, Ochitani, & Peters 1991). These are high prices to pay when only small parts of the source documents may be relevant. The other possibility is to first reduce the incoming information to only the relevant parts before translating them into the analyst's language. This is useful, of course, only when the template fills are sufficiently reliable. This IE-MT order of combination is also likely to benefit from the fact that corresponding template data structures can be isomorphic across languages and that template slot fills considerably scale back from full-blown texts in their complexity.[2]

Figure 3 shows an analyst performing the IE task in a multilingual setting. Given a set of source documents in foreign languages, $L_d^1 ... L_d^n$, analysts who know language L_a (which may or may not be one of the source document languages) should be able to define their information requirements in L_a, and have the system summarize in L_a the result of IE from the source documents. The linguistic barrier must be crossed at two places in this setting:

1. the analyst's target information expressed in L_a must be translated into equivalent expressions in source document languages, $L_d^1 ... L_d^n$ [MT1]
2. the template data objects representing the extracted information in source languages must be turned into equivalent template data objects in the analyst language L_a [MT2]

These translations are much more constrained than translations of full source documents, so are likely to achieve high accuracy with a focused application of the existing MT technology.

Both IE and the information retrieval (IR) technology, which has also developed in the TIPSTER program, would be useful for information reduction. The

[2] Aone et al. (Aone *et al.* 1994) also report an assessment that the IE-MT order of combining the two technologies is more effective than the other, MT-IE order.

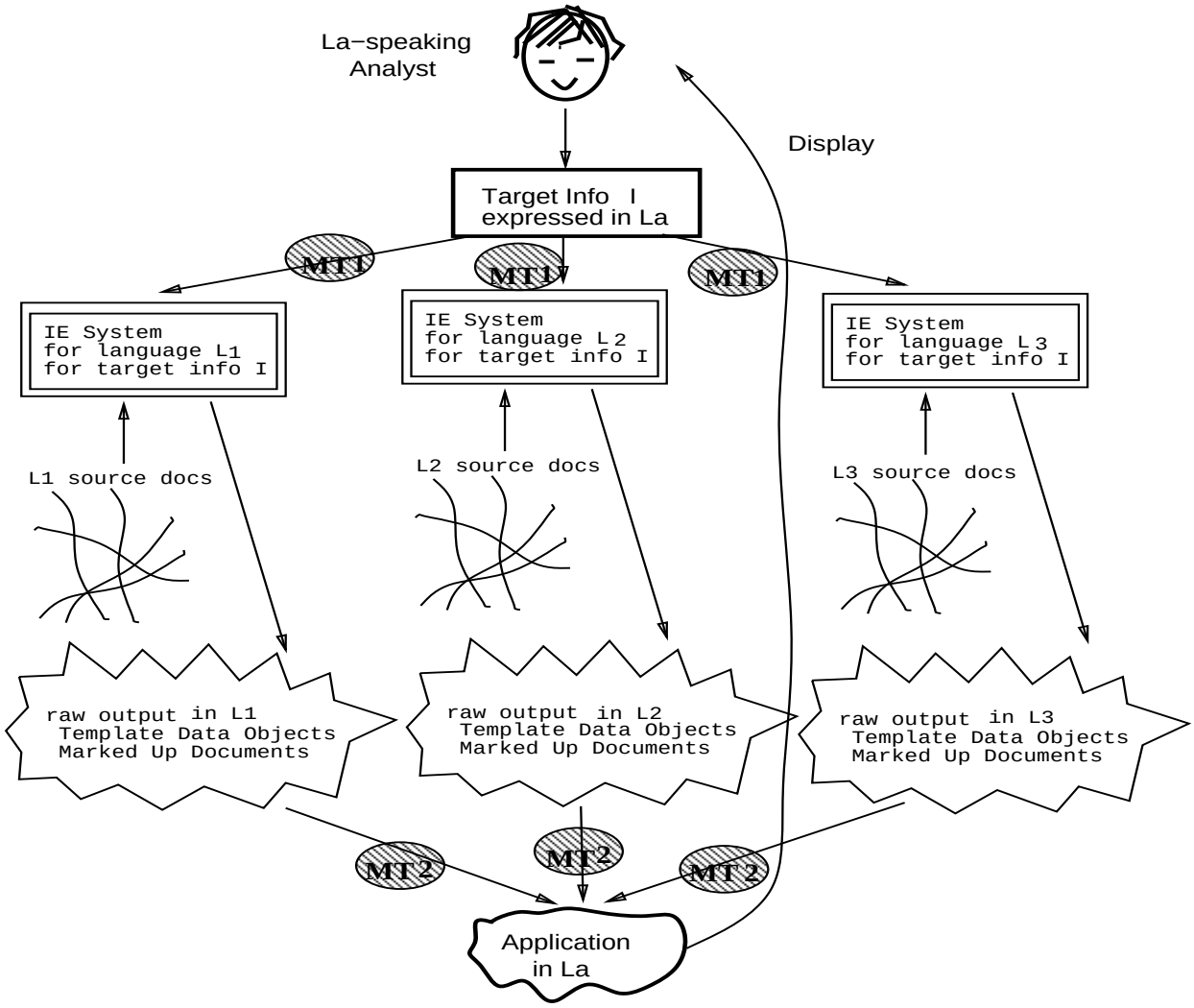

Figure 3: Multilingual Information Extraction

main functional difference between IE and IR is that IE outputs the specific *content* of the target information in a document, with the determination of relevance as a side effect, whereas IR retrieves the whole documents judged to be relevant. Since template data objects are easier to translate than full source documents, IE has a potential advantage over IR in multilingual information analysis.[3]

In sum, a multilingual IE system depicted in Figure 3 will inherit the robustness and speed of monolingual IE systems, and simplify the translation task by only focusing on template slot fills.

A key question in multilingual IE is how much of a monolingual IE infrastructure (e.g., rules and data structures) can be either reused or shared when more languages are added. This, however, can be answered only after a fully functioning bilingual IE system is in place.

Bilingual Information Extraction

We can demonstrate the key components of a full-blown multilingual IE system depicted in Figure 3 with a bilingual IE system that enables, for instance, an English-speaking analyst to obtain information from documents in a single non-English language. The purpose of this bilingual IE system is threefold:

1. to demonstrate that linguistic barriers can be overcome within an IE infrastructure
2. to demonstrate that the IE context allows scaling back from full machine translation
3. to identify the monolingual and bilingual components required for adding another language to the pool of source documents

A bilingual IE system consists of the following three components:

- a translation component that translates the analyst's description of the target information type into the corresponding descriptions in the source document language [MT1]
- an open-target monolingual IE system that outputs template data structures for given target information [IE]
- a template translation component that turns the source language templates into target language templates [MT2]

[3]Since IR can handle larger volumes of information more efficiently, it may be good to first do IR, then do IE only on those documents determined to be relevant by IR. This also follows from an experimental result (Bear, Israel, & Kehler 1996), where IE's content extraction results improved IR's relevance ordering of documents for a common target information.

Japanese-English Named Entity Translation

As an initial demonstration of a bilingual IE system, we are in the process of extending FASTUS (Appelt *et al.* 1993b)(Appelt *et al.* 1993a)(Appelt *et al.* 1995)(Hobbs *et al.* 1992)(Hobbs *et al.* 1996b) into a Japanese named-entity extraction system for English-speaking analysts. See Figure 4. This Japanese-English named entity extraction system extracts named entities (i.e., organization, person, location, date, time, money, and percent) from Japanese (and English) documents and present them in English. We are developing a prototype component that automatically translates Japanese named entity templates into English templates, taking advantage of available on-line corpora to develop and evaluate the prototype system. The following are examples of output template objects after translation.

```
<DOC-1>
  DATE: 19910611
  SOURCE: Nikkei
  LANGUAGE: Japanese
  ENTITIES: <ORG-1><ORG-2><PERS-1><PERS-2>
  ENTITY-RELS: <PERS-ORG-REL-1><PERS-ORG-REL-2>

<ORG-1>
  NAME: ``MITI''
  TYPE: GOVERNMENT
  COUNTRY: JAPAN

<ORG-2>
  NAME: ``Industrial Policy Bureau''
  TYPE: GOVERNMENT
  COUNTRY: JAPAN

<PERS-1>
  NAME: ``Nakao''
  NATIONALITY: JAPAN

<PERS-2>
  NAME: ``Yuji Tanabashi''
  NATIONALITY: JAPAN

<PERS-ORG-REL-1>
  PERSON: <PERS-1>
  ORG: <ORG-1>
  POSITION: ``minister''
  STATUS: CURRENT

<PERS-ORG-REL-2>
  PERSON: <PERS-2>
  ORG: <ORG-2>
  POSITION: ``chief''
  STATUS: CURRENT
```

Named entity extraction benefits from IE's well-established name recognition capability. Name translation, on the other hand, is one of the weaknesses of

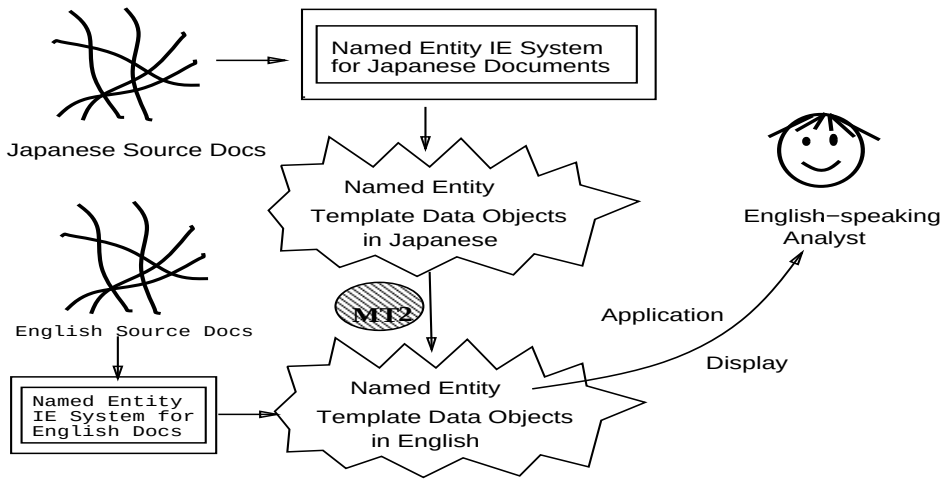

Figure 4: Japanese-English Named Entity Extraction

current MT systems — even the best commercial MT systems performed poorly in recognizing and translating names (Steve Maiorano, personal communication). This IE-MT approach to named entities, therefore, could also become a useful special-purpose component in a full-scale MT system.

In this system, the analyst's query is implicitly assumed to be something like "Report all the named organizations, named persons, named locations, and specific dates and times." So the MT1 translation is unnecessary. The system thus consists of two major components:

1. Japanese FASTUS as a monolingual IE system [IE]
2. a Japanese-English named entity translation component [MT2]

These components are described in more detail in this section.

FASTUS

Japanese FASTUS's basic architecture, shown in Figure 5, is unchanged from the English FASTUS (Hobbs *et al.* 1996b). The input document is first tokenized, then ASCII characters are sent to the ASCII Tokenizer, and 2-byte characters are sent to JUMAN, a public-domain Japanese morphological analyzer from Kyoto University. The ASCII Tokenizer is identical to the English FASTUS Tokenizer, which recognizes alphabetic, alphanumeric, numeric, and separator tokens as well as SGML tag tokens. JUMAN analyzes the input Japanese string, which lacks spaces between words, into a single best sequence of morphemes. These morphemes are turned into FASTUS Lexical Item objects with slots for literal string, normalized string, lexical category, inflection type, and so forth.

The mixed sequence of ASCII and morpheme tokens is then input into the SGML Handler, which recognizes the document structure based on SGML tags, and outputs a FASTUS Document object with slots for the headline, text, and other SGML fields. The headline slot has a sequence of sentences. The text slot has a sequence of paragraphs, each of which contains a sequence of sentences. The SGML Handler is written in a declarative pattern specification language (called FASTSPEC), so it can be easily adapted to non-SGML text tagging formats, as well as to more complex text structures containing sections, subsections, and tables.

The Document object is input into the Sentence Loop consisting of a sequence of finite-state transducers, namely, the Preprocessor, Name Recognizer, Basic Phrase Recognizer, Complex Phrase Recognizer, and Clause-level Event Recognizer. These linguistic phases recognize increasingly complex expressions in the sentence, recording syntactic and semantic attributes and producing template objects. At the end of each sentence loop, the Merger merges the new and existing template objects produced from the document so far. Document processing outputs a set of template objects that represent extracted information.

Japanese FASTUS produces a template entity for each Organization, Person, or Location name and Date, Time, Money, and Percent expression. Most of the names are recognized in the Name Recognizer phase based on internal patterns. After the Name Recognizer phase, the Alias Recognition routine recognizes some names or unknown words as aliases of longer names recognized earlier in the same document. The Parser and Combiner phases recognize a name's surrounding linguistic contexts, sometimes converting a phrase of one type into a phrase of another type.

We will also add a Coreference Resolution module that resolves reference of anaphoric mentions of Organizations, Persons, and Locations so that additional properties of these entities can be added to the data structures.

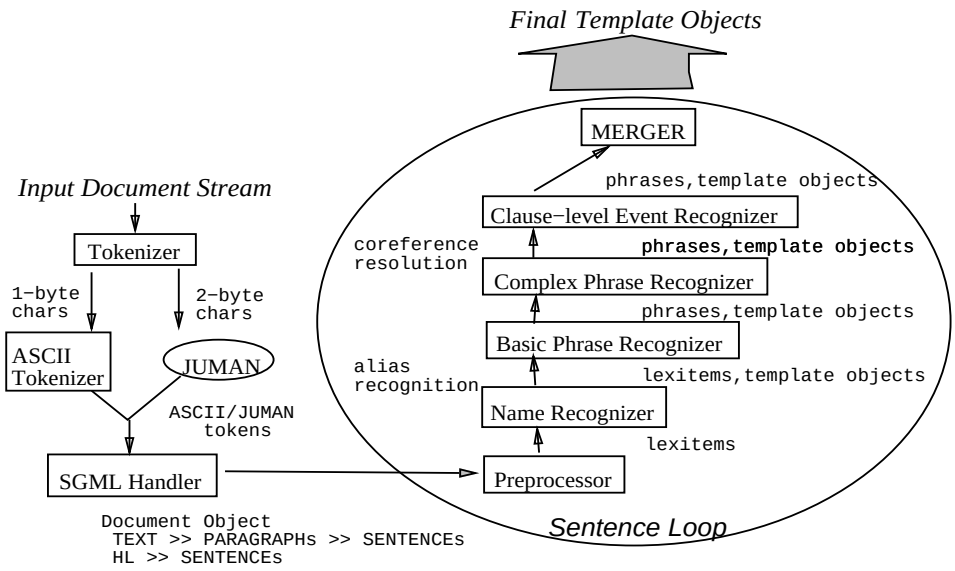

Figure 5: Japanese FASTUS System Overview

Named Entity Translation

The Japanese template output is then translated into corresponding English templates. Template translation involves recursive translation of the basic template structure and slot values. The basic template structures are assumed to be isomorphic between Japanese and English. Types of slot values are other template objects, set fills, or string fills. Set fill values are members of predefined sets of atoms — for instance, organization subtypes are, roughly, COMPANY, GOVERNMENT, or OTHER — so they can be in a one-to-one mapping across languages (e.g., COMPANY $\Leftrightarrow$ KIGYOU, GOVERNMENT $\Leftrightarrow$ SEIHU). String fills thus remain as the only nontrivial problem in template translation.

String fills are names, words, or arbitrary noun phrases. In Japanese-English translation of name strings, we need to solve the following three problems:

Kanji Names Produce the Roman alphabet versions of Japanese or Chinese proper names given in kanji characters.

Katakana Names Produce the original spellings of foreign proper names given in katakana characters.

Organization Names Produce the official English versions of multiword organization names.

For kanji-name alphabetization, we need a list of common last names, first names, and location names in Japanese and Chinese with their pronunciations (in hiragana). Katakana names pose a special difficulty because Japanese phonology tends to wipe out certain distinctions in original languages (e.g., r for both l and r, ch for both t and ch). It is desirable, then, to use a list of common katakana names with their original spellings.

The hardest and open-ended problem is with translation of organization names because it cannot simply be a compositional translation of words in the name string. In the JPRS report of 1,540 organization names (Foreign Broadcast Information Service (FBIS) 1995), we find about equal proportions of three types of translations between Japanese and English: (1) compositional translations without reconfiguration, (2) compositional translations with reconfiguration, and (3) noncompositional (or idiosyncratic) translations that insert or omit name parts. Some examples are shown here with indices for corresponding parts. (Original Japanese names are in kanji characters.)

1. Compositional without Reconfiguration
   - J: denryoku$_1$ kenkyuu$_2$ jo$_3$
     E: Electric$_1$ Power$_1$ Research$_2$ Institute$_3$
   - J: chou-dendou$_1$ sangyou$_2$ gijutsu$_3$ kaihatsu$_4$ kondan$_5$ kai$_6$
     E: Superconductor$_1$ Industry$_2$ Technology$_3$ Development$_4$ Consultation$_5$ Group$_6$
2. Compositional with Reconfiguration
   - J: denki$_1$ tsushin$_1$ sinkou$_2$ kai$_3$
     E: Association$_3$ for the Promotion$_2$ of Telecommunications$_1$
   - J: denki$_1$ zetsuen-butsu$_2$ shori$_3$ kyoukai$_4$
     E: Association$_4$ for the Treatment$_3$ of Electric$_1$ Insulator$_2$
3. Idiosyncratic Insertion or Omission
   - J: ajia$_1$ denki$_2$ tsushin$_2$ gijutsu$_3$ kyouryoku$_4$ zaidan$_5$

(*lit.:* Asia$_1$ Telecommunication$_2$ Technology$_3$ Cooperation$_4$ Foundation$_5$)
E: Asia$_1$ Teleteco$_{2,3,4}$ Organization$_5$

- J: denkii$_1$ tsushin$_1$ seisaku$_2$ sougou$_3$ kenkyu$_4$ jo$_5$
(*lit.:* Telecommunications$_1$ Policy$_2$ General$_3$ Research$_4$ Institute$_5$)
E: Research$_4$ Institute$_5$ of Telecom-Policies$_{1,2}$ **and Economics**
- J: denpa$_1$ gijutsu$_2$ kyoukai$_3$
(*lit.:* Radio$_1$ Technology$_2$ Association$_3$)
E: Radio$_1$ **Engineering and Electronics** Association$_3$
- J: chikyu$_1$ kansoku$_2$ eisei$_3$ chousei$_4$ kaigi$_5$
(*lit.:* Earth$_1$ Observation$_2$ Satellites$_3$ **Adjustment**$_4$ Committee$_5$)
E: Committee$_5$ of Earth$_1$ Observation$_2$ Satellites$_3$
- J: chikyu$_1$ sangyou$_2$ bunka$_3$ kenkyu$_4$ jo$_5$
(*lit.:* Earth$_1$ Industry$_2$ Culture$_3$ Research$_4$ Institute$_5$)
E: Global$_1$ Industrial$_2$ **and Social Progress** Research$_4$ Institute$_5$
- J: chikyu$_1$ kankyou$_2$ sangyou$_3$ gijutsu$_4$ kenkyu$_5$ kikou$_6$
(*lit.:* Earth$_1$ **Environment**$_2$ **Industry**$_3$ Technology$_4$ Research$_5$ Organization$_6$)
E: Research$_5$ Institute$_6$ of **Innovative** Technologies$_4$ for the Earth$_1$

The first two kinds can be handled by general-purpose name translation *rules*, but the last kind must make use of a known translation list. It is thus crucial to collect available resources such as bilingual lists of government organizations and major companies. As a start, we wil make use of the JPRS Report (Foreign Broadcast Information Service (FBIS) 1995), which contains about 1,540 government organization names in Japanese with their English translations.

We will establish an evaluation method for this template translation component. To substantiate our claim that the IE-MT order is more effective than the MT-IE order, we might compare the output of the IE-MT system with the output of an MT-IE system, which, for instance, first uses a commercial MT system to translate a set of Japanese articles into English, then extracts named entities from them with the English FASTUS.

Example-based Name Translation

The initial rule-based name translation component described above is a good starting point under limited time and resources, but if there is a sufficiently large bilingual name list, it is feasible to develop a corpus-based name translation. *Example-based* machine translation (e.g., (Sato 1991)) is a corpus-based approach suitable for translating tightly constrained structured expressions. Name expressions are an excellent candidate for this.

	Monolingual	Bilingual	Multilingual
Closed Target	completed (MUC-4,5,6)	under dev. (FASTUS)	future
Open Target	under dev. (FASTUS)	future	future

Table 1: Information Extraction: Where are We Now?

An example-based name translation will make use of three additional on-line resources — a large Japanese-English bilingual name list, a Japanese-English dictionary, and a Japanese thesaurus. When a new organization name is encountered, the system looks for the most *similar* name in the example database. The conceptual distances in the thesaurus are used as a measure of similarity here. The example database contains a list of known translation pairs, with links between corresponding parts of each pair. The system then composes the best translation for the new name from the known or plausible translations of its parts. For instance, suppose that the above list of translation pairs are in the database and the new example is *chikyu sangyou bunka kondan kai* (*lit.*: Earth Industry Culture Consultation Group). Its translation is composed by putting together the known translation of *chikyu sangyou bunka* as *Global Industrial and Social Progress* and the known translation of *kondan kai* as *Consultation Group*, resulting in *Global Industrial and Social Progress Consultation Group*. When the example database is larger, there may be more examples of literally translating *chikyu sangyou bunka* into *Global Industry and Culture*, in which case, the preferred translation would also take this form.

An example-based name translation approach requires a large bilingual name list. Since existing bilingual name resources may be limited in scope, we plan to add an example acquisition component. The system will output newly found names and their translation hypotheses after each run, and add them to the permanent list after human verification and correction. As more new documents are processed, more items will be added to the permanent resource. There are a number of Japanese news corpora to be mined for this purpose.

Conclusion

As the information extraction technology matures, we can be more ambitious about its actual utility for information analysts. We have mapped out a path from the current closed-target monolingual IE technology toward the eventual open-target multilingual IE technology. See Table 1.

As a logical next step along the path, we have described a bilingual named entity translation system under development. This scaled back system can demonstrate the feasibility of one of the crucial breakthroughs necessary along the path, namely, the fact that transla-

tion of template slot values avoids some of the hardest problems in machine translation. This named entity translation system also promises to be effective as a component of a full-blown MT system by providing a focused solution to one of the weaknesses of the current MT systems.

Acknowledgment

I would like to thank Steve Maiorano for helpful discussions.

References

Aone, C.; Blejer, H.; Okurowski, M. E.; and Ess-Dykema, C. V. 1994. A hybrid approach to multilingual text processing: Information extraction and machine translation. In *Technology Partnerships for Crossing the Language Barrier (Proceedings of the First Conference of the Association for Machine Translation in the Americas (AMTA))*, 1–7.

Appelt, D.; Hobbs, J.; Bear, J.; Israel, D.; Kameyama, M.; and Tyson, M. 1993a. SRI: description of the JV-FASTUS system used for MUC-5. In *Proceedings of the 5th Message Understanding Conference*, 221–236. DARPA.

Appelt, D.; Hobbs, J.; Bear, J.; Israel, D.; and Tyson, M. 1993b. FASTUS: a finite-state processor for information extraction from real-world text. In *Proceedings of the International Joint Conference on Artificial Intelligence*.

Appelt, D.; Hobbs, J.; Bear, J.; Israel, D.; Kameyama, M.; Kehler, A.; Martin, D.; Myers, K.; and Tyson, M. 1995. SRI International FASTUS system: MUC-6 test results and analysis. In *Proceedings of the 6th Message Understanding Conference*, 237–248. DARPA.

Bear, J.; Israel, D.; and Kehler, A. 1996. Using information extraction to improve document retrieval. SRI International Artificial Intelligence Center.

Foreign Broadcast Information Service (FBIS). 1995. Science & Technology (Japan): Japanese organizations and their english translations. Technical Report JPRS-JST-95-044, Foreign Broadcast Information Service (FBIS), Washington, DC.

Grishman, R. 1995. The NYU system for MUC-6 or where's the syntax? In *Proceedings of the 6th Message Understanding Conference*, 167–176. DARPA.

Hobbs, J. R., and Kameyama, M. 1990. Translation by abduction. In *Proceedings of the Seventh ICCL, COLING-90*.

Hobbs, J.; Appelt, D.; Bear, J.; Israel, D.; and Tyson, M. 1992. FASTUS: a system for extracting information from natural-language text. Technical Report Technical Note No. 519, SRI International Artificial Intelligence Center.

Hobbs, J.; Appelt, D.; Bear, J.; Israel, D.; Kameyama, M.; Stickel, M.; and Tyson, M. 1996a. SRI's Tipster II project. In *Proceedings of TIPSTER 24-month Conference*. DARPA. To appear.

Hobbs, J. R.; Appelt, D. E.; Bear, J.; Israel, D.; Kameyama, M.; Stickel, M.; and Tyson, M. 1996b. FASTUS: A cascaded finite-state transducer for extracting information from natural-language text. In Roche, E., and Schabes, Y., eds., *Finite State Devices for Natural Language Processing*. MIT Press, Cambridge,Massachusetts.

Kameyama, M., and Arima, I. 1993. A minimalist approach to information extraction from spoken dialogues. In *Proceedings of the International Symposium on Spoken Dialogue (ISSD-93)*, 137–140. Tokyo, Japan: ISSD Organizing Committee, Waseda University.

Kameyama, M., and Arima, I. 1994. Coping with aboutness complexity in information extraction from spoken dialogues. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP-94)*, 87–90.

Kameyama, M.; Kawai, G.; and Arima, I. 1996. A real-time system for extracting information from human-human spontaneous spoken dialogues. In *Proc. International Conference on Spoken Language Processing (ICSLP-96)*.

Kameyama, M.; Ochitani, R.; and Peters, S. 1991. Resolving translation mismatches with information flow. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, 193–200. Berkeley, Calif.: Association for Computational Linguistics.

Kameyama, M. 1995. Information extraction from spontaneous spoken dialogues. SRI International Artificial Intelligence Center.

Kameyama, M. 1996. MET name recognition with Japanese FASTUS. In *Proceedings of TIPSTER 24-month Conference*. DARPA. To appear.

Sato, S. 1991. *Example-based Machine Translation*. Ph.D. Dissertation, Kyoto University.

Sundheim, B. 1995. Overview of results of the MUC-6 evaluation. In *Proceedings of the 6th Message Understanding Conference*, 13–32. DARPA.