Semantics

WORD SENSE ACQUISITION FOR MULTILINGUAL TEXT INTERPRETATION *

Paul S. Jacobs
Information Technology Laboratory
GE Research and Development Center
Schenectady, NY 12301 USA
psjacobs@crd.ge.com

Abstract

We discuss a method for using automated corpus analysis to acquire word sense information for multilingual text interpretation. Our system, SHOGUN, extracts data from news stories with broad coverage in Japanese and English. Our approach focuses on tying together word senses, using a combination of world knowledge (ontology) with word knowledge (corpus data). We explain the approach and its results in SHOGUN.

1. INTRODUCTION

Text interpretation research has recently come to focus on data extraction – the problem of producing structured information from free text, usually to populate a database. Once the key information has been extracted, it can be used to help analyze the contents of large volumes of texts, detect trends, and retrieve selected information. Data extraction is at the center of the problem of managing large volumes of text.

Our group has led data extraction work for a number of years, developing new architectures and lexicons for natural language processing and testing these methods in a variety of applications [Jacobs and Rau, 1993; Jacobs, 1990; Jacobs and Rau, 1990]. In the last two years, as part of the U.S. government's ARPA TIPSTER program, we have extended this research to handle broader domains, with higher accuracy, and to process texts in multiple languages [Jacobs *et al.*, 1993].

The goal of processing texts in a new language is not only to show that the basic algorithms are language-independent, but also to preserve as much *knowledge* as possible across languages, and, where applicable, across domains. For example, in adapting an English system to handle Japanese texts, it is important that the Japanese system configuration makes use as much as possible of the general knowledge, and even the English vocabulary, that the system has. This maximizes the performance, and minimizes the amount of work, each time the system is applied to a new language.

*This research was sponsored in part by the Advanced Research Projects Agency (DOD) and other government agencies. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Advanced Research Projects Agency or the US Government.

SHOGUN is unique in a number of ways, but it is particularly distinguished by the sharing of knowledge resources in different languages. The approach to multilingual interpretation involves two key elements: First, the system includes a core ontology of about 1,000 concepts that support word senses in the core English and Japanese lexicons, which are also identical in structure. Second, our system acquires much of its domain-specific knowledge, including combinations of words and phrases, from corpus data, easing the mapping of word class information into a new language. For example, the English verb *establish* corresponds very closely to the Japanese word *setsuritsu* (設立). In the TIPSTER domain of joint ventures, both *establish* and *setsuritsu* are used to describe the creation of companies ("establish a joint venture"), products ("establish a telecommunications and data network"), facilities ("establish a factory"), and other more abstract concepts (e.g. "establish a stronger foothold in Europe"). The TIPSTER task, which requires distinct information for companies, facilities, activities, and products, makes it crucial to distinguish these different word usages – regardless of language.

SHOGUN's results on the final TIPSTER benchmark compared very favorably to those of other systems [Jacobs *et al.*, 1993]. There are many different ways to view and analyze the many different benchmark statistics, but the area in which SHOGUN's approach was most clearly distinguished was in *recall* – the percentage of data from each test set that was correctly extracted by the program. On this measure, SHOGUN extracted, on average, 37% more correct information than any other system in any configuration. SHOGUN had somewhat lower *precision* (13% lower on average) than the highest precision system in each configuration, meaning that SHOGUN also produced a somewhat larger amount of incorrect information than other systems. The system, in both languages, often identified information that was not found by any other system, a result that we attribute to having better coverage in its knowledge base than other systems.

The rest of this paper will describe the problem of multilingual interpretation as it appears in the TIPSTER task, then present our solution, emphasizing knowledge structures and knowledge acquisition.

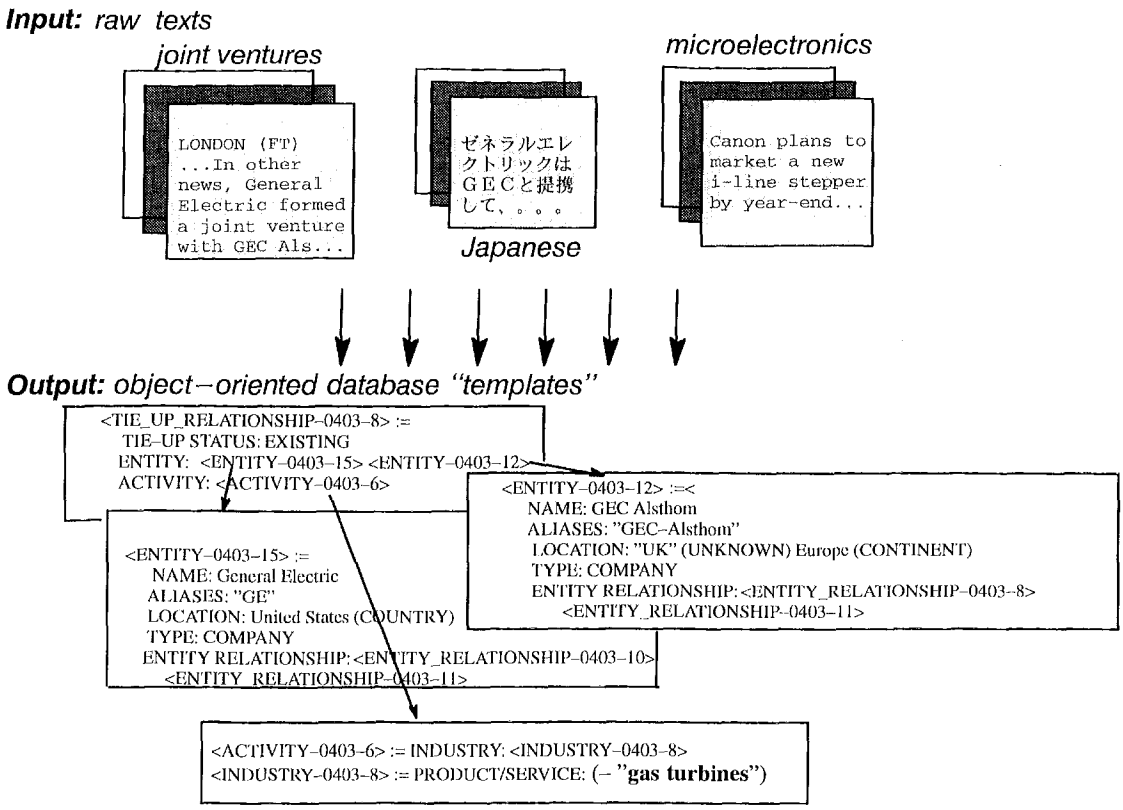

Figure 1: The TIPSTER (MUC-5) data extraction task

2. TIPSTER TASKS

TIPSTER is a program of the U.S. government Advanced Research Projects Agency (ARPA).** To emphasize portability across languages and domains, the teams in TIPSTER data extraction were required to develop capabilities and perform benchmark tests in two languages—English and Japanese—and two domains—microelectronics and joint ventures—resulting in four sets of benchmark results in each evaluation. The final evaluation, known as MUC-5 [Sundheim, 1993], was held in August, 1993, and included the four TIPSTER data extraction contractors as well as 13 other sites from four countries.

Figure 1 illustrates the basic TIPSTER data extraction task. In each configuration, systems process a set of texts and produce a set of database entries, or *templates*. The templates are specified as part of each domain; thus the Japanese templates in the joint venture domain are almost identical in structure to the English joint venture templates. The task, for each text, combines the recognition of high-level concepts (such as the identification of a joint venture in a text) with the discrimination of the meaning of individual phrases (such as descriptions of products) and the resolution of references. For example, Figure 2 shows a very simple example of a production joint venture between two companies.

For each of these texts, the data that must be extracted includes the generation of typed objects (such as entities and relationships) and slot fills that incorporate information, either directly or through inferences, from the texts. Much of this information comes from the recognition of high-level entities and relationships such as that shown in Figure 1. The rest includes much more detailed information, such as the activity, facilities and financing involved in a joint venture. Figure 3 shows this part of the information for the sample text, in the format of the actual correct responses, with italicized annotations to show where the information comes from in the example.

The slot fills in TIPSTER templates include "set fills"—drawn from a fixed list, such as the text code `PRODUCTION` for manufacturing and the numerical code `20` ("Food and kindred products")*** for processed food production, "string fills"—drawn from the ac-

**Our project, which included GE Corporate Research and Development, the Center for Machine Translation at Carnegie Mellon University, and Martin Marietta Management and Data Systems (formerly GE Aerospace), was one of four teams in the data extraction component of TIPSTER.

***The numerical codes for the `PRODUCT/SERVICE` slot (and the comparable 製品・サービス slot in Japanese) use the major groupings of the U. S. government Standard Industry Classification (SIC) scheme.

<DOCNO> 0659 </DOCNO>
<DD> SEPTEMBER 28, 1989, THURSDAY </DD>
<SO> Copyright (c) 1989 Kyodo News Service </SO>
<TXT>
KIKKOMAN CORP. WILL LINK UP WITH A TAIWANESE FOOD FIRM IN OCTOBER TO PRODUCE SOY SAUCE IN TAIWAN, COMPANY OFFICIALS SAID THURSDAY.

PRESIDENT KIKKOMAN, CAPITALIZED AT 80 MILLION TAIWAN YUAN (ABOUT 440 MILLION YEN), WILL BE OWNED 50 PERCENT EACH BY KIKKOMAN AND PRESIDENT ENTERPRISES CORP., TAIWAN'S LARGEST FOODSTUFF MAKER.

THE JOINT VENTURE WILL MANUFACTURE SOY SAUCE AT THE TAIWANESE FIRM'S PLANT WITH KIKKOMAN'S TECHNOLOGICAL ASSISTANCE AND DISTRIBUTE THE PRODUCT UNDER THE KIKKOMAN BRAND NAME. THE ANNUAL SALES TARGET IS SET AT AROUND 3,000 KILOLITERS WITHIN A FEW YEARS, THEY SAID.

....

Figure 2: A sample input text

<FACILITY-0659-1> :=
  LOCATION: Taiwan (COUNTRY)
  TYPE: FACTORY
<INDUSTRY-0659-1> :=
  INDUSTRY-TYPE: PRODUCTION
  PRODUCT/SERVICE: (20 "SOY [SAUCE]")
<INDUSTRY-0659-2> :=
  INDUSTRY-TYPE: SALES
  PRODUCT/SERVICE: (51 "SOY [SAUCE]") / (51 "[THE PRODUCT]")
<ACTIVITY-0659-1> :=
  INDUSTRY: <INDUSTRY-0659-1>
  ACTIVITY-SITE: (<FACILITY-0659-1> <ENTITY-0659-3>)
  START TIME: <TIME-0659-1>
<ACTIVITY-0659-2> :=
  INDUSTRY: <INDUSTRY-0659-2>
  ACTIVITY-SITE: (Taiwan (COUNTRY) <ENTITY-0659-3>)
<TIME-0659-1> :=
  DURING: 1989
<OWNERSHIP-0659-1> :=
  OWNED: <ENTITY-0659-3>
  TOTAL-CAPITALIZATION: 80000000 TWD
  OWNERSHIP-%: (<ENTITY-0659-1> 50) (<ENTITY-0659-2> 50)

...IN OCTOBER TO PRODUCE SOY SAUCE IN TAIWAN,

THE JOINT VENTURE WILL MANUFACTURE SOY SAUCE AT THE TAIWANESE FIRM'S PLANT

...AND DISTRIBUTE THE PRODUCT...

THE JOINT VENTURE WILL MANUFACTURE SOY SAUCE AT THE TAIWANESE FIRM'S PLANT

...IN OCTOBER TO PRODUCE SOY SAUCE IN TAIWAN,

...CAPITALIZED AT 80 MILLION TAIWAN YUAN...
WILL BE OWNED 50 PERCENT EACH BY KIKKOMAN AND PRESIDENT ENTERPRISES CORP.

Figure 3: Part of correct answer for text 0659

tual text, such as ''SOY SAUCE'', pointers to other objects — such as <ENTITY-0659-1> and a variety of "normalized" fills — such as Taiwan (COUNTRY). The set fills often capture local information in the text, while the objects (consisting of an identifier with a related template fills) often involve inferences from many different parts of the text. For example, in this case, the object ACTIVITY-0659-1 reflects the fairly subtle distinction that the venture will be manufacturing soy sauce at President Enterprises' plant (the result of reference resolution) but that the sales will be carried out somewhere else in Taiwan (the result of a real inference). In this part of the task, major object-level decisions often hinge on the interpretation of the individual words, making the task very lexicon-intensive.

Interpreting the activity information; that is, identifying what each venture is doing along with the appropriate products and codes, requires knowledge about word usage in context. Activity words like *build*, *establish*, and *create* are just as common as words like *produce* and *manufacture*. In many cases, whether something is a joint venture activity or not depends on a fairly detailed analysis of these words — "building a factory" is different from "building a new plane", "building a business", and of course, from "building a presence". These similar phrases can not only evoke different product codes, but also can often affect the high-level construal of a story. The interpretations of word senses come together with domain and task knowledge in extracting the appropriate information from these phrases.

Because one of the goals of this project was to develop methods of handling new domains and languages, it was important to cope with these crucial differences in word usage in a general way. This meant partitioning the knowledge of the system into four components: (1) generic, (2) domain-dependent, (3) language-dependent, and (4) domain-and-language-dependent. With the detail of analysis that parts of the task require, such as those described above, it is essential not only to minimize the amount of knowledge that is dependent on either language or domain, but also to minimize the effort of acquiring knowledge that is dependent on either domain or language, and, especially, knowledge that is dependent on both. The sections that follow will cover these aspects of our so-

lution to the TIPSTER problem.

3. LEXICON & ONTOLOGY

The previous section framed some of the problems of data extraction in TIPSTER, with an emphasis on the aspects of the task that require substantial amounts of knowledge. We also presented our approach to the task by explaining the synergistic objectives of creating generic resources and developing knowledge acquisition methods. This section will focus on the generic resources, while the next section will concentrate on acquisition methods.

The main generic resource of SHOGUN is its core ontology of about 1,000 concepts, which was developed to support GE's NLToolset lexicon [Jacobs and Rau, 1993; McRoy, 1992] and had been tested fairly thoroughly on a variety of data extraction tasks prior to TIPSTER. We augmented the core ontology using the CMU ontology from machine translation [KBM, 1989] and used the extended ontology as the basis for Japanese lexicon development. The idea of this effort was that the Japanese lexicon would mirror the existing English lexicon, allowing for sharing of the domain-independent components of the knowledge base across languages as well as the sharing of any domain-specific knowledge that would be added.

For example, the following is the English entry for the verb *establish* and its related forms:

```
( establish
   :POS verb
   :G-DERIVS ((-er noun tr_actor) (-ment noun ...))
   :SENSES
   (( establish1
      :EXAMPLES (she established superiority * ... )
      :SYNTAX (one-obj thatcomp whcomp prespart)
      :TYPE *primary*
      :PAR (c-causal-event)
      :SYNONYMS (set_up)    )
    ( establish2
      :EXAMPLES (the court established fault)
      :SYNTAX (one-obj thatcomp whcomp prespart)
      :TYPE *primary*
      :PAR (c-deciding)
      :SYNONYMS (determine)
    ))
   :X-DERIVS
   (( establish-ment-x
      :X-DERIVS (-ment noun)
      :EXAMPLES (the eating establishment)
      :EXPRESS c-organization
    ))  )
```

The Japanese lexicon now consists of about 13,000 words. This is somewhat more than the 10,000 unique roots of the English lexicon, but the English lexicon is still much richer in morphology and more thoroughly tested than the Japanese. Nevertheless, the two lexicons are roughly comparable and certainly compatible. For example, the Japanese entry for *setsuritsu* (設立) is the following:

```
( 設立
:POS nsa
:G-DERIVS ()
:SENSES
(( setsuritsu1
:SYNTAX ()
:EXAMPLES (合弁会社を設立する)
:TYPE *primary*
:PAR (c-causal-event)
:SYNONYMS (establish set_up)
:NOTE (:nttd-kana ("せつりつ") :jv-dom)
)) )
```

The main link between the English and Japanese lexicons is through the `:PAR` field (for *parent*) in each word sense, which joins that sense to its parent in the ontology. In this case, the common parent between *establish* and *setsuritsu*, `c-causal-event` (the bringing about of events or effects), is a fairly general category that includes two senses of *open* as well as a variety of others like *duplicate* and *bridge*. The reason that *establish* ends up in this general class is that it is very hard to confine any sense of the word to creation events.

Having a shared ontology and lexicon format has certain advantages. It is a requirement for using a common language processing framework across languages, and it ensures that words with similar meanings in different languages end up with similar representations and ontological restrictions. The next section discusses how this common framework must be extended for domain-specific usage.

4. ACQUISITION

In a task like TIPSTER, we cannot capture all the subtle distinctions that the task requires in the core lexicon. Each domain, like joint ventures, requires a large amount of very specific knowledge, not only about how words like *establish* behave, but also about simple facts like that *office supplies* usually includes things like pens and papers while *office equipment* usually includes machines like computers and copiers. Because many of these facts are at the intersection of world knowledge and word knowledge (that is, they are patterns of language use that reflect real-word concepts), even the most specific pieces of knowledge often seem to apply across languages.

The degree to which ontology contributes to interpretation in any particular domain was, in general, somewhat less than we might have expected. For example, the category `c-causal-event` includes not only words that don't have anything to do with joint ventures, but also words that in the joint venture domain could be misinterpreted. The category in Japanese includes senses of words like 新設 and 併設, which behave very similarly to *setsuritsu* (設立), but doesn't include many others that also behave similarly. In English joint ventures, the extended class of words used to describe the establishment of a new company includes *plan*, *set up*, *form*, and *create*. In Japanese, the class includes 設け, 施設, 新設, 設立, つく, and

設置. In both cases, these word classes are determined from examining corpus data, with a particular emphasis on words that are used to describe the formation of new companies. This includes words from different ontological groups and excludes certain words from the `c-causal-event` category.

As we have pointed out, words like *establish* and *setsuritsu* are so critical to the understanding of joint ventures that knowledge about such words can be hand coded for each language and domain. However, doing this hand-coding for many aspects of the TIPSTER task would not only involve an extraordinary amount of effort, but it would thwart one of the main objectives of the project—to develop methods that ease portability across languages and domains.

Our "middle ground" solution to capturing the more specialized knowledge, relying neither on generic knowledge nor on language-specific encodings, was to create word classes to represent the information needed in the TIPSTER data extraction task, to apply these word classes across languages, and to expand them using automated corpus analysis. We observed that, although Japanese and English had different vocabularies and properties, the usage of words in each Japanese corpus was very similar to the usage of comparable English words in corpora from the same domain. For example, the word *equipment* in English joint ventures is very similar to the word *souchi* (装置) in Japanese, and the task-specific distinctions are the same in English and Japanese (e.g., the distinctions among medical equipment, transportation equipment, and electrical equipment).

We took advantage of this observation in developing a two-stage process of developing word groupings across languages. Once the major groupings were defined (manually), the automated process of corpus analysis consisted of (1) expanding word classes by associating common, relatively unambiguous words with other classes, and (2) further expanding and identifying ambiguities using a "bootstrapping" process. The bootstrapping process used the knowledge that had already been encoded to classify a chunk of text (for example, deciding that a particular phrase described transportation equipment), and assuming that words with a high degree of association with that category must also be related.

The first stage of the process started with, for both English and Japanese, a set of words that were closely identified with business activities (like "manufactures", and "distributes"). Using a corpus of about 10 million words (English from the *Wall Street Journal* and Japanese from *Nikkei Shinbun*), we took the words that were most likely to appear within a window of three words of an "activity" word, and tried, manually, to assign them to product classes. The statistical analysis used a weighted mutual information statistic. This resulted in initial groupings of words into classes corresponding to particular product groups, or codes. For example, the following is the English class corresponding roughly to SIC code 38, "Measuring, analyzing, and controlling instruments":

```
biomedical copier copiers lens lenses
instrument pacemakers photocopy photocopier
photocopiers radar navigational microfilm
monitoring navigation guidance avionics photo
photographic photography camera clocks watches
eyeglasses sunglasses glasses Polaroid frames
```

The second stage of corpus analysis was the "bootstrapping" process. From the texts that included the "good" activity terms, the program assigned a set of word classes, such as that above, based on its existing knowledge base. For example, if "eyeglasses" appeared in an activity text, that text would be assigned to group 38, along with whatever other categories also appeared. Then, for each word appearing in every text in the corpus, we again applied the mutual information statistic to find the significant relationships between words and groups. When a word could be associated with more than one group, this process identified phrases that could help to distinguish the word sense, and collected such ambiguous words in a separate list so that they could be dealt with manually, if necessary.

Figure 4, shows, for a Japanese sample, the results of the corpus analysis process, including the identification of the "product" words, with frequencies and weights, and the analysis of whether the corpus data confirmed what was known about each word.

In the TIPSTER benchmarks, we relied on manually-corrected lists, using the statistical weights only to help resolve differences in selecting among multiple potential product descriptions. However, in our own tests, we found the performance of the manually edited knowledge on the activity portion of the template to be only slightly better than the fully automated sample. The knowledge base of word groups included over 4000 words in English and over 2000 in Japanese.

Although SHOGUN has been tested in a series of government benchmarks, we still consider this method to be only a starting point. There are many problems. The corpora used for training were not a good representative sample, because they were drawn from different sources from the test samples due to limitations in the availability of representative training materials. The Japanese training relied on segmenting the training corpus into words, a process that occasionally introduced error. Other sources of error included cases where our initial manual groupings involved misinterpretations of the task.

Nevertheless, both the core ontology and the automated training method had a significant impact on SHOGUN's results in TIPSTER. The next section presents a brief summary of these results.

5. RESULTS

Figure 5 shows the overall recall, precision and F-measure scores for SHOGUN on the four configuration of the final TIPSTER (MUC-5) benchmark. EJV and

#	Score	Word	> (find-industry-conflicts)
424	52.1	ソフト	メディア ambiguous: ((36 3475 4.6) (48 390 4.1) (73 1082 4.7) (78 205 5.9) (27 62 4.5))
373	77.3	製品	機体 new: ((45 47 4.8))
327	88.9	よる	学術 ambiguous: ((28 1994 6.8) (20 400 4.6))
314	95.2	写真	専管 ambiguous: ((62 128 5.0) (63 160 6.1) (65 80 5.1))
288	38.2	向け	リーダー new: ((26 120 9.0))
242	36.2	装置	センタ ambiguous: ((36 5640 4.4) (48 1026 4.6))
218	47.3	国内	L S I ambiguous: ((36 3977 4.4) (48 913 5.0))
198	29.2	コンピューター	再生 new: ((26 5 5.8))
189	57.8	現地	総菜 new: ((54 2 4.1))
186	51.4	部品	カード confirmed: ((61 1201 5.6))
154	28.5	カード	石油 new: ((13 274 4.4))
137	33.1	半導体	ミリオン new: ((61 26 5.2))
128	25.8	ビル	仕入れ new: ((54 5 4.1))
127	352.3	新工場	インターコネクティビティ ambiguous: ((36 1800 6.5) (48 144 5.5) (73 552 6.5) (78 108 7.7))
125	93.0	リゾート	キヤノン new: ((78 36 4.1))
123	26.2	車	コンポーネント ambiguous: ((36 6603 6.2) (73 2026 6.3) (48 528 5.2) (78 396 7.5))
118	988.6	P O S	中古車 new: ((33 36 4.0))
118	37.0	ブランド	不動産 confirmed: ((65 656 5.1))
110	108.1	ソフトウエア	半導体 ambiguous: ((36 11231 4.3) (73 3437 4.3) (78 677 5.5))
108	26.1	通信	エンジン confirmed: ((37 1174 4.4))
105	32.9	不動産	ロボット confirmed: ((35 264 4.2))
104	37.5	力	ガソリン confirmed: ((13 48 4.7))
93	58.1	石油	アイスクリーム different: ((54 4 4.1)) vs. (20)
			.....

Figure 4: Some results of corpus analysis

JJV are the English and Japanese joint venture tests, and EME and JME are the two microelectronics test sets. Recall is the percentage of possible information that is correctly identified by the system. Precision is the percentage of information produced by the system that is correct. The F-measure is the geometric mean of recall and precision.

	Rec	Pre	F-meas
EJV	57	49	52.8
JJV	57	64	60.1
EME	50	48	49.2
JME	60	53	56.3

Figure 5: SHOGUN Scores for MUC-5

Scores as low as 50 recall may appear low, and certainly leave room for improvement. A 50 recall measure means that the system only correctly recovered half of the possible information, on average, from each text. However, by a number of relative comparisons, these numbers are good. They are a significant improvement over previous benchmarks, and are close to the recall and precision scores of the GE system on much easier tests. The TIPSTER task is quite difficult, with trained human intelligence analysts often producing recall scores in the 70s.

As we have pointed out, SHOGUN's recall was, on average, 37% higher than any other system in each configuration, although the precision was 13% lower than the system with the best precision in each configuration. For example, the next best system in English joint ventures (EJV) had 38 recall and 58 precision, and the next best system in Japanese joint ventures (a different system) had 42 recall and 67 precision.

Much of the difference in performance between SHOGUN and other systems can be attributed to difficult portions of the task, where SHOGUN sometimes had recall scores as much as 3 or 4 times as high as other systems. The portions of the joint venture template shown in Figure 3 are examples of such components. Because these were the most knowledge-intensive components of the task, we believe that the results validate SHOGUN's approach to knowledge acquisition. Certainly the system had much better coverage than other systems, and we attribute this result to the representation and automation used in word sense interpretation.

Figure 6 gives an informal analysis of the level and type of effort used in each configuration. Although the Japanese scores were generally higher than English, the Japanese configurations largely relied on the English knowledge development. The level of effort for Japanese joint ventures was higher than English because the English system started out with much more than the Japanese system (for example, we already had a fairly well developed English name recognition component). By contrast, the Japanese microelectronics configuration derived almost entirely from the English, with almost no effort required from Japanese speakers.

Many other sites participated in the TIPSTER project and the MUC evaluations, including two others [Cowie and Pustejovsky, 1993; Weischedel *et al.*, 1993] that covered both domains and both languages, and one other [Lehnert *et al.*, 1993] that focused on lexical acquisition, although only in English. In addition,

Domain/Language	Effort/Skill Level	Other Notes
English joint ventures	1 person–year, system developers, native English speakers	Some effort not reflected in results Difficult to measure because of many experiments
Japanese joint ventures	1.5 person–years, mostly Japanese college students with non–native developers	Least efficient, but most interesting effort Best overall results
English micro–electronics	3 person–months, system developers, native speakers, no knowledge of ME	Lowest overall results (but explained by sample variation)
Japanese micro–electronics	2 person–months, non–developers, non–native speakers (with some help from natives, developers)	Last configuration done, least work, good results (but not refined)

Figure 6: Effort required for each domain and language

there has been other significant related work in robust processing of texts, notably [Hobbs *et al.*, 1992]; however, this research has generally emphasized syntactic coverage rather than lexical coverage. Finally, related research in lexical acquisition [Zernik, 1991] focuses on core lexical resources rather than on customizing the lexicon through the use of a representative corpus. Hence, the research that we have presented has advanced the state of the art both in the use of the corpus to identify word sense information and the demonstration of multilingual capabilities.

6. CONCLUSION

SHOGUN's approach to word sense interpretation across languages uses language-independent information to group word senses in different languages. A core ontology of 1,000 concepts links senses that are domain-independent. However, in the tasks on which SHOGUN has been tested, domain-specific word sense information is more critical. For this more specialized sense knowledge, the system uses an innovative method of training and bootstrapping using a corpus, using a statistical analysis to help assign words and phrases to language-independent groupings. This approach significantly sped the acquisition of word sense information in TIPSTER, resulting in high coverage on the most difficult components of the task.

References

[Cowie and Pustejovsky, 1993] J. Cowie and J. Pustejovsky. Description of the DIDEROT system as used for TIPSTER text. In *Proceedings of the TIPSTER Phase I Final Meeting*, September 1993.

[Hobbs *et al.*, 1992] J. R. Hobbs, D. E. Appelt, J. Bear, M. Tyson, and D. Magerman. Robust processing of real-world natural-language texts. In Paul S. Jacobs, editor, *Text-Based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval.* Lawrence Erlbaum Associates, Hillsdale, NJ, 1992.

[Jacobs and Rau, 1990] Paul Jacobs and Lisa Rau. SCISOR: Extracting information from on-line news. *Communications of the Association for Computing Machinery*, 33(11):88–97, November 1990.

[Jacobs and Rau, 1993] P. S. Jacobs and L. F. Rau. Innovations in text interpretation. *Artificial Intelligence*, 63:143–191, 1993.

[Jacobs *et al.*, 1993] P. Jacobs, G. Krupka, L. Rau, M. Mauldin, T. Mitamura, T. Kitani, I. Sider, and L. Childs. The TIPSTER/SHOGUN project. In *Proceedings of the TIPSTER Phase I Final Meeting*, San Mateo, CA, September 1993. Morgan Kaufmann.

[Jacobs, 1990] Paul Jacobs. To parse or not to parse: Relation-driven text skimming. In *Proceedings of the Thirteenth International Conference on Computational Linguistics*, pages 194–198, Helsinki, Finland, 1990.

[KBM, 1989] The KBMT Report. Technical report, Center for Machine Translation, Carnegie Mellon University, 1989.

[Lehnert *et al.*, 1993] W. Lehnert, J. McCarthy, S. Soderland, E. Riloff, C. Cardie, J. Peterson, and F. Feng. Description of the CIRCUS system used for TIPSTER text extraction. In *Proceedings of the TIPSTER Phase I Final Meeting*, September 1993.

[McRoy, 1992] Susan McRoy. Using multiple knowledge sources for word sense discrimination. *Computational Linguistics*, 18(1), March 1992.

[Sundheim, 1993] Beth Sundheim, editor. *Proceedings of the Fifth Message Understanding Conference (MUC-5).* Morgan Kaufmann Publishers, San Mateo, Ca, August 1993.

[Weischedel *et al.*, 1993] R. Weischedel, D. Ayuso, S. Boisen, H. Fox, R. Ingria, T. Matsukawa, C. Papageorgiou, D. MacLaughlin, M. Kitagawa, T. Sakai, J. Abe, H. Hosihi, Y. Miyamoto, and S. Miller. BBN PLUM executive summary. In *Proceedings of the TIPSTER Phase I Final Meeting*, September 1993.

[Zernik, 1991] U. Zernik, editor. *Lexical Acquisition: Using On-Line Resources to Build a Lexicon.* Lawrence Erlbaum Associates, Hillsdale, NJ, 1991.