Cross-lingual Information Extraction System Evaluation

Kiyoshi Sudo, Satoshi Sekine, and Ralph Grishman
Department of Computer Science
New York University
715 Broadway, 7th Floor,
New York, NY 10003
{sudo,sekine,grishman}@cs.nyu.edu

Abstract

In this paper, we discuss the performance of cross-lingual information extraction systems employing an automatic pattern acquisition module. This module, which creates extraction patterns starting from a user's narrative task description, allows rapid customization to new extraction tasks. We compare two approaches: (1) acquiring patterns in the source language, performing source language extraction, and then translating the resulting templates to the target language, and (2) translating the texts and performing pattern discovery and extraction in the target language. We demonstrate an average of 8-10% more recall using the first approach. We discuss some of the problems with machine translation and their effect on pattern discovery which lead to this difference in performance.

1 Introduction

Research in information extraction (IE) and its related fields has led to a wide range of applications in many domains. The portability issue of IE systems across different domains, however, remains a serious challenge. This problem is being addressed through automatic knowledge acquisition methods, such as unsupervised learning for domain-specific lexicons (Lin et al., 2003) and extraction patterns (Yangarber, 2003), which require the user to provide only a small set of lexical items of the target classes or extraction patterns for the target domain. The idea of a self-customizing IE system emerged recently with the improvement of pattern acquisition techniques (Sudo et al., 2003b), where the IE system customizes itself across domains given by the user's query.

Furthermore, there are demands for access to information in languages different from the user's own. However, it is more challenging to provide an IE system where the target language (here, English) is different from the source language (here, Japanese): a cross-lingual information extraction (CLIE) system.

In this research, we explore various methods for efficient automatic pattern acquisition for the CLIE system, including the translation of the entire source document set into the target language. To achieve efficiency, the resulting CLIE system should (1) provide a reasonable level of extraction performance (both accuracy and coverage) and (2) require little or no knowledge on the user's part of the source language. Today, there are basic linguistics tools available for many major languages. We show how we can take advantage of the tools available for the source language to boost extraction performance.

The rest of this paper is organized as follows. Section 2 and 3 discuss the self-adaptive CLIE system we assess throughout the paper. In Section 4, we show the experimental result for entity detection. Section 5 discusses the problems in translation that affect the pattern acquisition and Section 6 discusses related work. Finally, we conclude the paper in Section 7 with future work.

2 Query-Driven Information Extraction

One approach to IE portability is to have a system that takes the description of the event type from the user as input and acquires extraction patterns for the given scenario. Throughout the paper, we call this kind of IE system QDIE (Query-Driven Information Extraction) system, whose typical procedure is illustrated in Figure 1.

QDIE (e.g. (Sudo et al., 2003a)) consists of three phases to learn extraction patterns from the source documents for a scenario specified by the user.

First, it applies morphological analysis, dependency parsing and Named Entity (NE) tagging to the entire source document set, and converts all the sentences in the source document set into dependency trees. The NE tagging replaces named entities by their class, so the resulting dependency trees contain some NE class names as leaf nodes. This is crucial to identifying common patterns, and to applying these patterns to new text.

Second, the user provides a set of narrative sen-

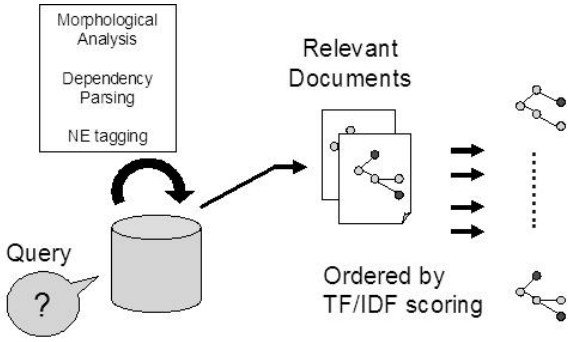

Figure 1: QDIE Pattern Acquisition

tences describing the scenario (the events of interest). Using these sentences as a retrieval query, the information retrieval component of QDIE retrieves representative documents of the scenario specified by the user (*relevant documents*).

Then from among all the possible connected subtrees of all the sentences in the *relevant documents*, the system calculates the score for each pattern candidate. The scoring function is based on TF/IDF scoring in IR literature; a pattern is more relevant when it appears more in the *relevant documents* and less across the entire collection of source documents. The final output is the ordered list of pattern candidates.

Note that a pattern candidate contains at least one NE, so that it can be used to match a portion of a sentence which contains an instance of the same NE type. The matched NE instance is then extracted. The pattern candidates may be simple predicate-argument structures (e.g. (resign from {C-POST}) in business domain) or even a complicated subtree of a sentence which commonly appears in the *relevant documents* (e.g. ({C-ORG} report personnel affair (that {C-PERSON} resigns))).

3 Cross-lingual Information Extraction

(Riloff et al., 2002) present several approaches to cross-lingual information extraction (CLIE). They describe the use of "cross-language projection" for CLIE, exploiting the word alignment of documents in one language and the same documents translated into a different language by a machine translation (MT) system. They conducted experiments between two relatively close languages, English and French. In the experiment reported here, we will explore CLIE for two more disparate languages, English and Japanese.

The QDIE system can be used in a cross-lingual setting, and thus, the resulting cross-lingual version of the QDIE system can minimize the requirement of the user's knowing the source language. Figure 2 shows two possible ways to achieve this goal.

It may be realized by translating all the documents of the source language into the target language, and then running the monolingual version of the QDIE system for the target language (Translation-based QDIE). In our experiment, we translated all the source Japanese documents into English. Then we ran English-QDIE system to get the extraction patterns, which are used to extract the entities by pattern matching.

On the other hand, one can first translate the scenario description into the source language and use it for the monolingual QDIE system for the source language, assuming that we have access to the tools for pattern acquisition in the source language. Each entity in the extracted table is translated into the target language (Crosslingual-QDIE). In Figure 2, we implemented this procedure by first translating the English query into Japanese. [1] Then we ran Japanese-QDIE system to identify Japanese extraction patterns. The extraction patterns are used to extract items to fill the Japanese table. Finally, each item in the extracted table is separately translated into English. Note that translating names is easier than translating the whole sentences.

As we shall demonstrate, the errors introduced by the MT system impose a significant cost in extraction performance both in accuracy and coverage of the target event. However, if basic linguistic analysis tools are available for the source language, it is possible to boost CLIE performance by learning patterns in the source language. In the next section, we describe an experiment which compares these two approaches. In the following section, we assess the difficulty of learning extraction patterns from the translated source language document set caused by the errors of the MT system and/or the differences of grammatical structure of the translated sentences. We address specifically:

1. The accuracy of NE tagging on MT-ed source documents and the use of cross-language projection.
2. How the structural difference in source and target language affects the extracted patterns.
3. The reduced frequency of the extracted patterns, which makes it difficult for any measurement of pattern relevance to distinguish the

[1] Note that our current implementation uses the output from query translation by the MT system. As we note in Section 7, we plan to investigate the possibility of additional performance gain by using current crosslingual information retrieval techniques.

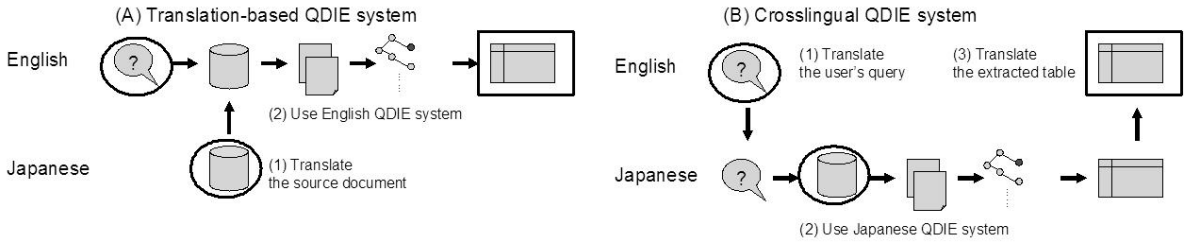

Figure 2: Translation-based QDIE System(A) vs Crosslingual QDIE System(B): The user's query (English), the source document (Japanese) and the target extracted table (English) are highlighted.

effective patterns of low frequency from the noise patterns.

4 Experiments

To evaluate the relevance of extraction patterns automatically learned for CLIE, we conducted experiments for the Translation-based QDIE system and the Cross-lingual QDIE system on the *entity extraction task*, which is to identify all the entities participating in relevant events in a given set of Japanese texts.

4.1 Experimental Setting

Since general NE taggers either are trained on English sentences or use manually created rules for English sentences, the deterioration of NE tagger's performance cannot be avoided if it is applied to the MT-ed English sentences. This causes the Translation-based QDIE system to identify fewer pattern candidates from the *relevant documents* since a pattern candidate must contain at least one of the NE types.

To remedy this problem, we incorporated "cross-language projection" (Riloff et al., 2002) only for Named Entities. We used word alignment obtained by using Giza++ (Och and Ney, 2003) to get names in the English translation from names in the original Japanese sentences. Note that it is extremely difficult to make an alignment of case markers where one language explicitly renders a marker as a word and the other does not. So, direct application of (Riloff et al., 2002) is not suitable for this experiment.

We compare the following three systems in this experiment.

1. Crosslingual QDIE system
2. Translation-based QDIE system with word alignment
3. Translation-based QDIE system without word alignment

4.2 Data

The scenario for this experiment is the Management Succession scenario of MUC-6(muc, 1995), where corporate managers assumed and/or left their posts. We used a much simpler template structure than the one used in MUC-6, with Person, Organization, and Post slots. To assess system performance, we measure the accuracy of the system at identifying the participating entities in a management succession event. This task does not involve grouping entities associated with the same event into a single template, in order to avoid possible effects of merging failure on extraction performance for entities.

The source document set from which the extraction patterns are learned consists of 132,996 Yomiuri Newspaper articles from 1998. For our Crosslingual QDIE system, all the documents are morphologically analyzed by JUMAN (Kurohashi, 1997) and converted into dependency trees by KNP (Kurohashi and Nagao, 1994). For the Translation-based QDIE system, all the documents are translated into English by a commercial machine translation system (IBM "King of Translation"), and converted into dependency trees by a corpus-based parser. We retrieved 1500 documents as *relevant documents*.

We accumulated the test set of documents by a simple keyword search. The test set consists of 100 Yomiuri Newspaper articles from 1999, out of which only 61 articles contain at least one management succession event. Note that all NE in the test documents both in the original Japanese and in the translated English sentences were identified manually, so that the task can measure only how well extraction patterns can distinguish the participating entities from the entities that are not related to any succession events. Table 1 shows the details of the test data.

4.3 Results

Each pattern acquisition system outputs a list of the pattern candidates ordered by the ranking function. The resulting performance is shown as a precision-

Documents (relevant + irrelevant)	100 $(61 + 39)$
Names (relevant + irrelevant)	Person: 173 + 651 Org: 111 + 709 Post: 210 + 626

Table 1: Statistics of Test Data

recall graph for each subset of top-n ranked patterns where n ranges from 1 to the number of pattern candidates. The parameters for each system are tuned to maximize the performance on separate validation data.

The association of NE classes in the matched patterns and slots in the template is made automatically; *Person*, *Organization*, *Post* (slots) correspond to C-PERSON, C-ORG, C-POST (NE-classes), respectively, in the Management Succession scenario.

Figure 3 shows the precision-recall curve for the top 1000 patterns acquired by each system on the entity extraction task. Crosslingual QDIE system reaches a maximum recall of 60%, which is significantly better than Translation-based QDIE with word alignment (52%) and Translation-based QDIE without word alignment (41%). Within the high recall range, Crosslingual QDIE system generally had better precision at the same recall than Translation-based QDIE systems. At the low recall range ($< 20\%$), the performance is rather noisy.

Translation-based QDIE without word alignment performs similarly to Translation-based QDIE with word alignment up to its maximum recall (41%). Translation-based QDIE with word alignment reached 10% higher maximum recall (52%).

5 Problems in Translation

The detailed analysis of the result revealed the effect of several problems caused by the MT system. The current off-the-shelf MT system's output resulted in difficulty in using it as a source of extraction patterns. In this section we will discuss the types of differences between the source and target languages, and their effect on pattern discovery.

Lexical differences Abbreviations in the source language may not have their corresponding short form in the target language. For example, "Kei-Dan-Ren" is an abbreviation of "*Kei*zai *Dan*tai *Ren*go-kai" which is an organization whose English translation is "Japan Federation of Economic Organizations". Such abbreviations may not be listed in the dictionary of the MT system. In such cases, the literal translation of the abbreviation may be difficult to recognize as a name and is likely to be treated as a common noun phrase.

Structural differences Some phrases in the source language may have more than one relevant translation. Depending upon the context where a phrase appears, the MT system has to choose one among the possible translations. Moreover, the MT system may make a mistake, of course, and output an erroneous translation. This results in a diverse distribution of extraction patterns in the target language. Figure 4 shows an example of such a case. Suppose an extraction pattern (({C-POST}-*ni*) *shuninsuru*) appears 20 times in the original Japanese document set, out of which it may be translated 10 times as (be appointed (to ({C-POST}))), 5 times as (assume ({C-POST})), 3 times as (be inaugurated (as ({C-POST}))), and 2 times as an erroneous translation. Some of the lower frequency translated patterns will be ranked lower by the scoring function and so will be hard to distinguish from noise.

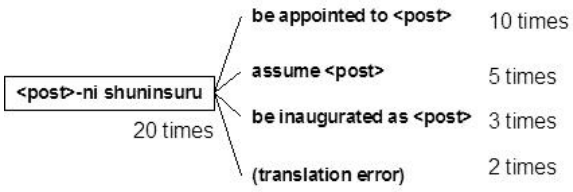

Figure 4: Example of Structural Difference in Translation: The translation of a Japanese expression into several English different expressions including erroneous ones.

Figure 5 shows an example of the case where the context around the name did not seem to be translated properly, so the dependency tree for the sentence was not correct. The right translation is "Okajima announced that President Hiroyuki Okajima, 40 years old, resigned formally ..." which results in the dependency between the main verb "announce" and the company "Okajima". The translation shown in Figure 5 not only shows incorrect word-translations, but also shows ungrammatical structure, including too many relative clauses. The structural error causes the errors in the dependency parse tree including having "end" as a root of the entire tree and the wrong dependency from "announced" to "the major department" in Figure 5 [2]. Thus, the accumulation of the errors resulted in missing the organization name "Okajima".

Also, the conjunctions in Japanese sentences could not be translated properly, and therefore, the

[2]The head is "the major department" and "announced" is modifying the head.

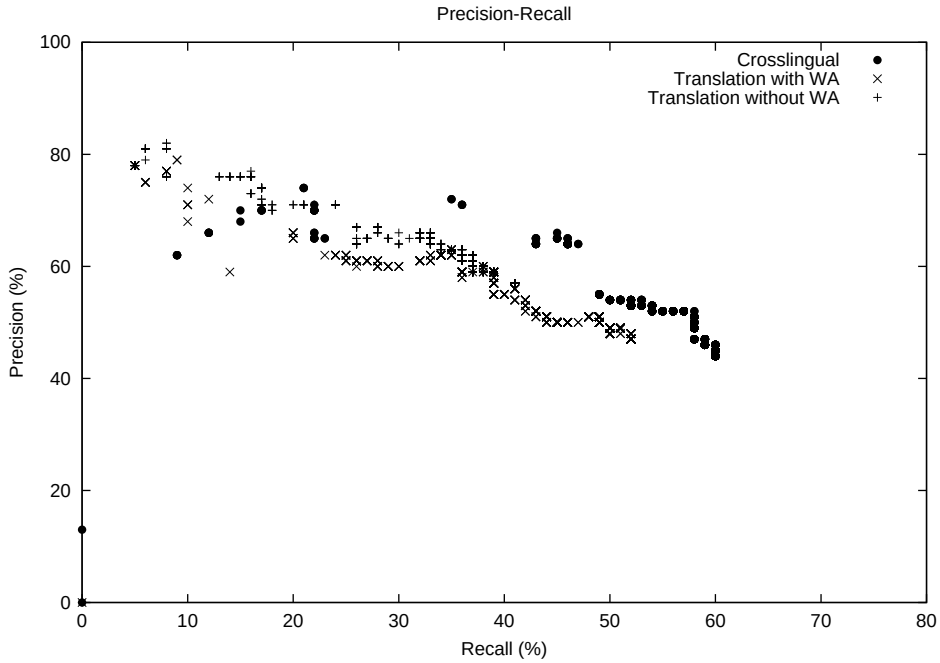

Figure 3: Performance Comparison on Entity Extraction Task

English dependency parser's output is significantly deteriorated. The example in Figure 6 shows the case where both "Mr. Suzuki" and "Mr. Asada" were inaugurated. In the original Japanese sentence, "Mr. Suzuki" is closer to the verb "be inaugurated". So, it seems that the MT system tries to find another verb for "Mr. Asada", and attaches it (incorrectly) to "unofficially arranged".

Out-of-Vocabulary Words The MT system may not have a word in the source language dictionary, in which case some MT systems output it in the original script in the source language. This happens not only for names but also for sentences which are erroneously segmented into words. Such problems, of course, may make it hard to detect Named Entities and get a correct dependency tree of the sentence.

However, translation of names is easier than translation of contexts; the MT system can output the transliteration of an unknown word. In fact, name translation of the MT system we used for this experiment is better than the sentence translation of the same MT system. The names appropriately extracted from Japanese documents by the Crosslingual QDIE system, in most cases, are correctly translated or transliterated if no equivalent translation exists.

6 Related Work

The work closest to ours is (Riloff et al., 2002). They showed how IE learning tools, bitext alignment, and an MT system can be combined to create CLIE systems between English and French. They evaluated a variety of methods, including one similar to our Translation-based QDIE. Their approaches were less reliant on language tools for the "source" language (in their case, French) than our Crosslingual-QDIE system. On the other hand, their tests were made on a closer language pair (English - French). We expect that the performance gap between Translation-based IE and Crosslingual IE is more pronounced with a more divergent language pair like Japanese and English.

There are interesting parallels between our work and that of (Douzidia and Lapalme, 2004), who discussed the role of machine translation in a crosslingual summarization system which produces an English summary from Arabic text. Their system took the same path as our Crosslingual QDIE: summarizing the Arabic text directly and only translating the summary, rather than translating the entire Arabic text and summarizing the translation. They had similar motivations: different translations produced by the MT system for the same word in different contexts, as well as translation errors, would interfere with the summarization process.

The trade-offs, however, are not the same for the two applications. For summarization either path requires an MT system which can translate entire sentences (either the original text or the summary). Translation-based QDIE has a similar requirement,

Output of MT system:

From Muika the term settlement of accounts ended February , 99 having become the prospect of the first deficit settlement of accounts after the war etc. , six of President Hiroyuki_Okajima (40) , two managing_directors , one managing_directors , the full-time_directors that are 13_persons submitted the resignation report , **"Okajima"** of Marunouchi , Kofu-shi who is the major department store within the prefecture announced that he resigns formally by the fixed general meeting of shareholders of the company planned at the end of this_month .

Output of Dependency Tree (part):

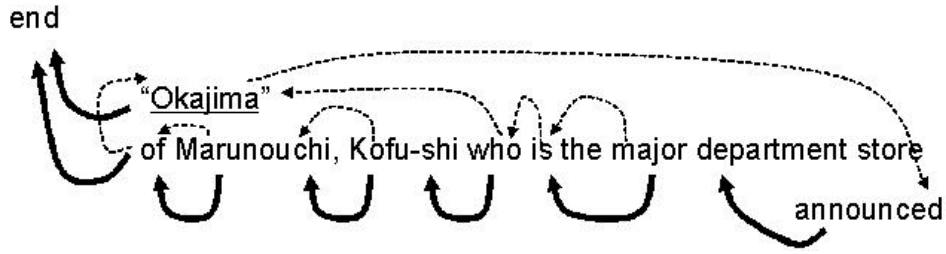

Figure 5: Example of Translation Errors: Figure also contains a part of the dependency parser's output of the sentence. Dashed lines show the correct dependencies.

but Crosslingual QDIE reduces the demands on MT: only query translation and name translation are required.

7 Conclusion

We discussed the difficulty in cross-lingual information extraction caused by the translation of the source documents using an MT system. The experimental result for entity extraction suggests that exploiting some basic tools available for the source language will boost the performance of the whole CLIE system.

We intend to investigate whether further performance gain may be obtained by introducing additional techniques for query translation. These techniques, including query translation on expanded queries and building a translation dictionary from parallel corpora, are currently used in crosslingual information retrieval (Larkey and Connell, 2003).

Acknowledgments

This research was supported in part by the Defense Advanced Research Projects Agency as part of the Translingual Information Detection, Extraction and Summarization (TIDES) program, under Grant N66001-001-1-8917 from the Space and Naval Warfare Systems Center, San Diego, and by the National Science Foundation under Grant ITS-00325657. This paper does not necessarily reflect the position of the U.S. Government.

References

Fouad Soufiane Douzidia and Guy Lapalme. 2004. Lakhas, an Arabic summarization system. In *Proceedings of DUC2004*.

Sadao Kurohashi and Makoto Nagao. 1994. KN parser : Japanese dependency/case structure analyzer. In *Proceedings of the Workshop on Sharable Natural Language Resources*.

Sadao Kurohashi, 1997. *Japanese Morphological Analyzing System: JUMAN*. http://www.kc.t.u-tokyo.ac.jp/nl-resource/juman-e.html.

Leah Larkey and Margaret Connell. 2003. Structured Queries, Language Modeling, and Relevance Modeling in Cross-Language Information Retrieval. In *Information Processing and Management, Special Issue on Cross Language Information Retrieval*.

Winston Lin, Roman Yangarber, and Ralph Grishman. 2003. Bootstrapped Learning of Semantic Classes from Positive and Negative Examples. In *Proceedings of the ICML-2003 Workshop on The Continuum from Labeled to Unlabeled Data*, Washington, D.C.

1995. *Proceedings of the Sixth Message Under-*

Output of MT system:

The personnel affairs to which managing_director Shosei_Suzuki (57) is inaugurated as the Nippon_Telegraph_and_Telephone president of the holding company which NTT will be the board of directors on the seventh , is inaugurated by NTT reorganization on July_1 at President Jun-ichiro_Miyatsu (63) the Nippon_Telegraph Telephone_East_Corporation (NTT_East_Japan) president of a local communication company , he is inaugurated as Vice_President Shuichi_Inoue (61) Nippon_Telegraph Telephone_West_Corporation (NTT western part of Japan) were unofficially arranged Vice_President Kazuo_Asada (59) the NTT_Communications president of a long distance international-telecommunications company .

Output of Dependency Tree (part):

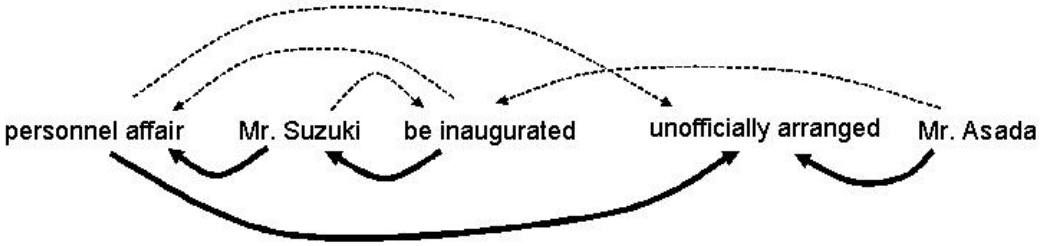

Figure 6: Example of Erroneous Conjunction Phrase: Figure also contains a part of the dependency parser's output of the sentence. Dashed lines show the correct dependencies.

standing Conference (MUC-6), Columbia, MD, November. Morgan Kaufmann.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Ellen Riloff, Charles Schafer, and David Yarowsky. 2002. Inducing Information Extraction Systems for New Languages via Cross-Language Projection. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*.

Kiyoshi Sudo, Satoshi Sekine, and Ralph Grishman. 2003a. An Improved Extraction Pattern Representation Model for Automatic IE Pattern Acquisition. In *Proceedings of the 41st Annual Meeting of Association of Computational Linguistics (ACL 2003)*, Sapporo, Japan.

Kiyoshi Sudo, Satoshi Sekine, and Ralph Grishman. 2003b. Pre-CODIE – Crosslingual On-Demand Information Extraction. In *Proceedings of HLT/NAACL 2003*, Edmonton, Canada.

Roman Yangarber. 2003. Counter-Training in Discovery of Semantic Patterns. In *Proceedings of the 41st Annual Meeting of Association of Computational Linguistics (ACL 2003)*, Sapporo, Japan.