Towards Translingual Information Access using Portable Information Extraction

Michael White, Claire Cardie, Chung-hye Han, Nari Kim,#
Benoit Lavoie, Martha Palmer, Owen Rambow,* Juntae Yoon

CoGenTex, Inc.
Ithaca, NY, USA
{mike,benoit,owen}
@cogentex.com

Institute for Research in
Cognitive Science
University of Pennsylvania
Philadelphia, PA, USA
chunghye@babel.ling.upenn.edu
{nari,mpalmer,jtyoon}
@linc.cis.upenn.edu

Dept. of Computer Science
Cornell University
Ithaca, NY, USA
cardie@cs.cornell.edu

Abstract

We report on a small study undertaken to demonstrate the feasibility of combining portable information extraction with MT in order to support translingual information access. After describing the proposed system's usage scenario and system design, we describe our investigation of transferring information extraction techniques developed for English to Korean. We conclude with a brief discussion of related MT issues we plan to investigate in future work.

1 Introduction

In this paper, we report on a small study undertaken to demonstrate the feasibility of combining portable information extraction with MT in order to support translingual information access. The goal of our proposed system is to better enable analysts to perform information filtering tasks on foreign language documents. This effort was funded by a SBIR Phase I award from the U.S. Army Research Lab, and will be pursued further under the DARPA TIDES initiative.

Information extraction (IE) systems are designed to extract specific types of information from natural language texts. In order to achieve acceptable accuracy, IE systems need to be tuned for a given topic domain. Since this domain tuning can be labor intensive, recent IE research has focused on developing learning algorithms for training IE system components (cf. Cardie, 1997, for a survey). To date, however, little work has been done on IE systems for languages other than English (though cf. MUC-5, 1994, and MUC-7, 1998, for Japanese IE systems); and, to our knowledge, none of the available techniques for the core task of learning information extraction patterns have been extended or evaluated for multilingual information extraction (though again cf. MUC-7, 1998, where the use of learning techniques for the IE subtasks of named entity recognition and coreference resolution are described).

Given this situation, the primary objective of our study was to demonstrate the feasibility of using *portable*—i.e., easily trainable—IE technology on Korean documents, focusing on techniques for learning information extraction patterns. Secondary objectives of the study were to elaborate the analyst scenario and system design.

2 Analyst Scenario

Figure 1 illustrates how an intelligence analyst might use the proposed system:

- The analyst selects one or more Korean documents in which to search for information (this step not shown).

Current affiliation: Konan Technology, Inc., Korea, nari@konantech.co.kr

* Current affiliation: ATT Labs-Research, Florham Park, NJ, USA, rambow@research.att.com

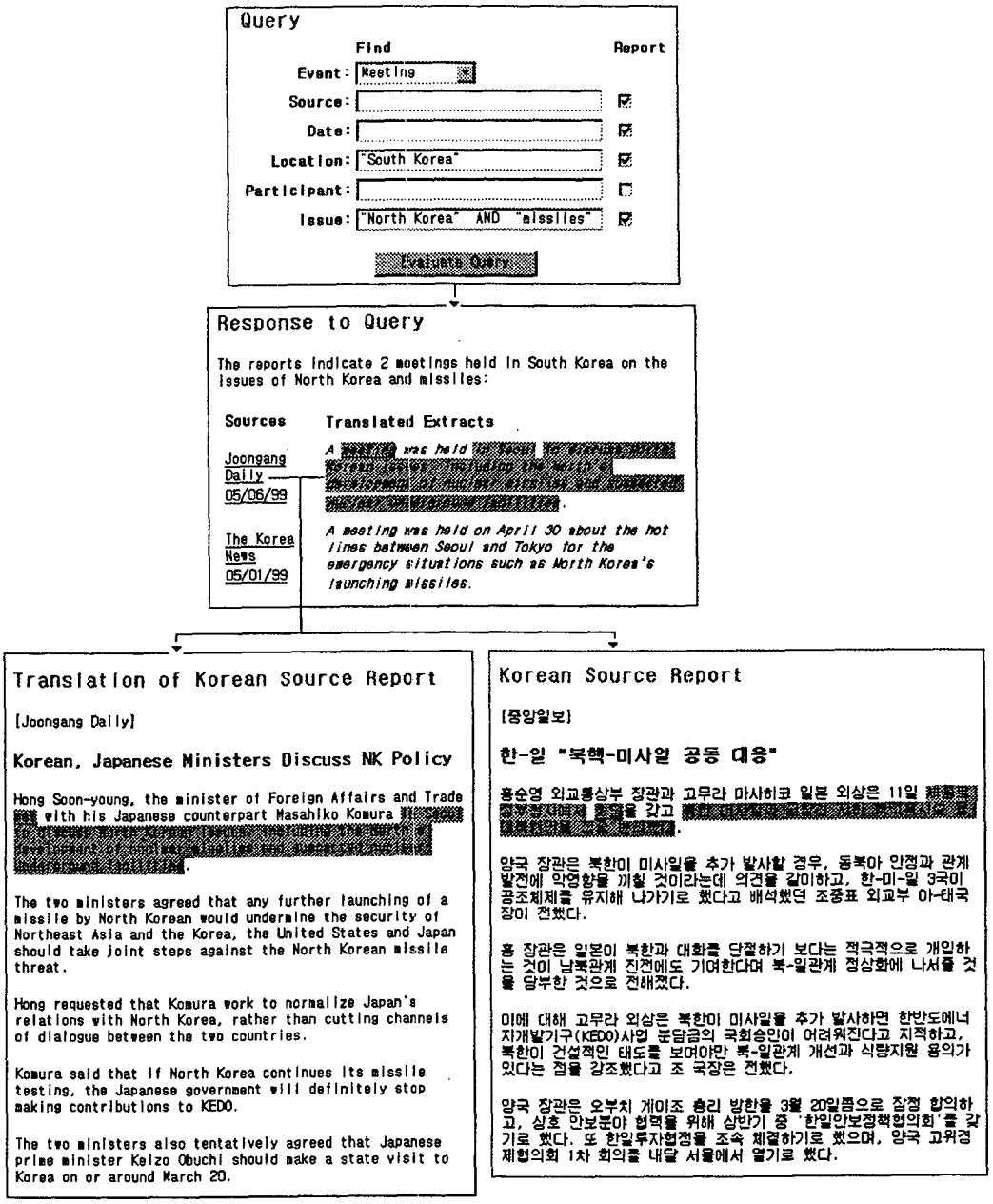

Figure 1

- The analyst selects one or more scenario template to activate in the query. Each scenario template corresponds to a specific type of event. Available scenario templates might include *troop movements*, *acts of violence*, *meetings and negotiations*, *protests*, etc. In Figure 1, the selected event is of type *meeting* (understood broadly).
- The analyst fills in the available slots of the selected scenario template in order to restrict the search to the information considered to be relevant. In Figure 1, the values specified in the scenario template indicate that the information to find is about *meetings* having as location *South Korea* and as issue *North Korea* and *missiles*. The analyst also

specifies what information s/he wants to be reported when information matching the query is found. In Figure 1, the selected boxes under the *Report* column indicate that all information found satisfying the query should be reported except for the *meeting participants*.[1]

- Once the analyst submits the query for evaluation, the system searches the input documents for information matching the query. As a result, a hypertext document is generated describing the information matching the query as well as the source of this information. Note that the query contains English keywords that are automatically translated into Korean prior to matching. The extracted information is presented in English after being translated from Korean. In Figure 1, the generated hypertext response indicates two documents in the input set that matched the query totally or in part. Each summary in the response includes just the translations of the extracted information that the analyst requested to be reported.
- For each document extract matching the analyst query, the analyst can obtain a complete machine translation of the Korean document where the match was found, and where the matched information is highlighted. Working with a human translator, the analyst can also verify the accuracy of the reported information by accessing the documents in their original language.

3 System Design

Figure 2 shows the high-level design of the system. It consists of the following components:

- The User Interface. The browser-based interface is for entering queries and displaying the resulting presentations.
- The Portable Information Extractor (PIE) component. The PIE component uses the Extraction Pattern Library — which contains the set of extraction patterns learned in the lab, one set per scenario template — to extract specific types of information from the input Korean documents, once parsed.
- The Ranker component. This component ranks the extracted information returned by the PIE component according to how well it matches the keyword restrictions in the query. The MT component's English-to-Korean Transfer Lexicon is used to map the English keywords to corresponding Korean ones. When the match falls below a user-configurable threshold, the extracted information is filtered out.
- The MT component. The MT component (cf. Lavoie et al., 2000) translates the extracted Korean phrases or sentences into corresponding English ones.
- The Presentation Generator component. This component generates well-organized, easy-to-read hypertext presentations by organizing and formatting the ranked extracted information. It uses existing NLG components, including the Exemplars text planning framework (White and Caldwell, 1998) and the RealPro syntactic realizer (Lavoie and Rambow, 1997).

In our feasibility study, the majority of the effort went towards developing the PIE component, described in the next section. This component was implemented in a general way, i.e. in a way that we would expect to work beyond the specific training/test corpus described below. In contrast, we only implemented initial versions of the User Interface, Ranker and Presentation Generator components, in order to demonstrate the system concept; that is, these initial versions were only intended to work with our training/test corpus, and will require considerable further development prior to reaching operational status. For the MT component, we used an early version of the lexical transfer–based system currently under development in an ongoing SBIR Phase II project (cf. Nasr et al., 1997; Palmer et al., 1998; Lavoie et al., 2000), though with a limited lexicon specifically for translating the slot fillers in our training/test corpus.

[1] While in this example the exclusion of participant information in the resulting report is rather artificial, in general a scenario template may contain many different types of information, not all of which are likely to interest an analyst at once.

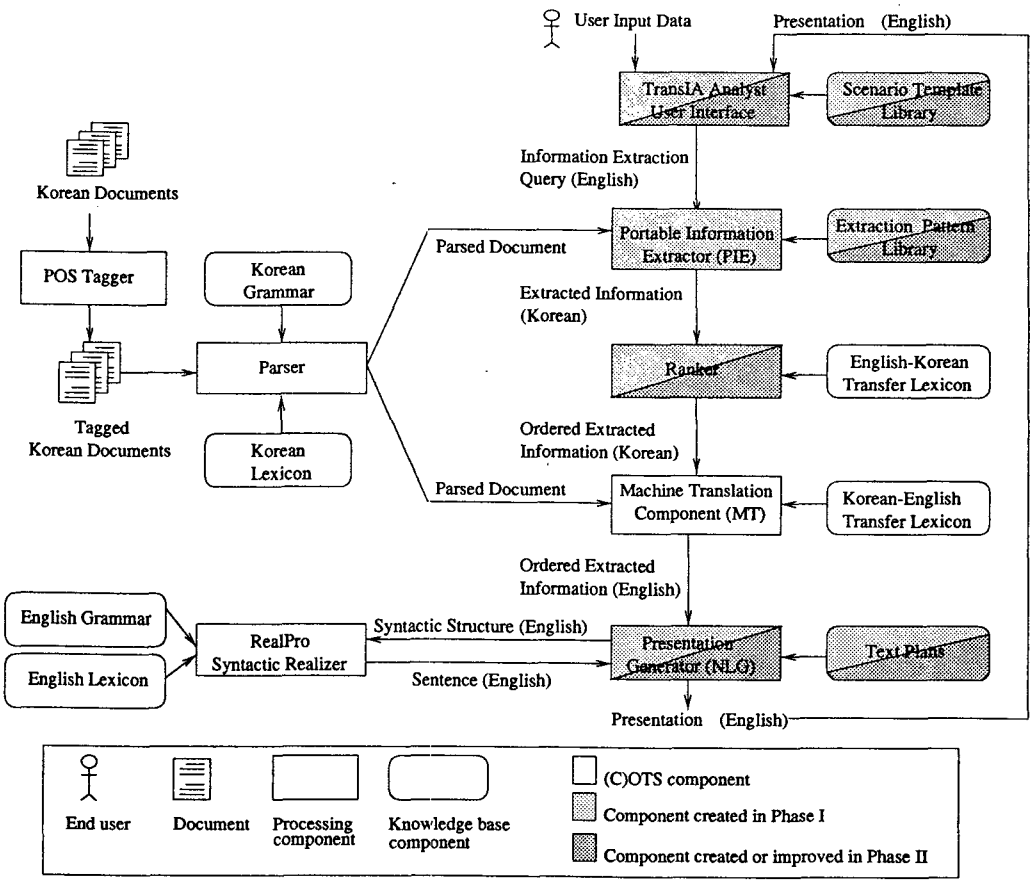

Figure 2

4 Portable Information Extraction

4.1 Scenario Template and Training/Test Corpus

For our Phase I feasibility demonstration, we chose a minimal scenario template for meeting and negotiation events consisting of one or more participant slots plus optional date and location slots.[2] We then gathered a small corpus of thirty articles by searching for articles containing "North Korea" and one or more of about 15 keywords. The first two sentences (with a few exceptions) were then annotated with the slots to be extracted, leading to a total of 51 sentences containing 47 scenario templates and 89 total correct slots. Note that in a couple of cases more than one template was given for a single long sentence.

When compared to the MUC scenario template task, our extraction task was considerably simpler, for the following reasons:

- The answer keys only contained information that could be found within a single sentence, i.e. the answer keys did not require merging information across sentences.
- The answer keys did not require anaphoric references to be resolved, and we did not deal with conjuncts separately.
- We did not attempt to normalize dates or remove appositives from NPs.

4.2 Extraction Pattern Learning

For our feasibility study, we chose to follow the AutoSlog (Lehnert et al., 1992; Riloff, 1993) approach to extraction pattern acquisition. In this approach, extraction patterns are acquired

[2] In the end, we did not use the 'issue' slot shown in Figure 1, as it contained more complex fillers than those that typically have been handled in IE systems.

```
1. E: <target-np>=<subject> <active voice verb>
      <participant> MET
   K: <target-np>=<subject> <active voice verb>
      <John-i>   MANNASSTA
      <John-nom> MET

2. E: <target-np>=<subject> <verb> <infinitive>
      <participant> agreed to MEET
   K: <target-np>=<subject> <verb1-ki-lo> <verb2>
      <John-un>  MANNA-ki-lo hapuyhayssta
      <John-nom> MEET-ki-lo agreed

   (-ki: nominalization ending, -lo: an adverbial postposition)
```

Figure 3

via a one-shot general-to-specific learning algorithm designed specifically for the information extraction task.[3] The learning algorithm is straightforward and depends only on the existence of a (partial) parser and a small set of general linguistic patterns that direct the creation of specific patterns. As a training corpus, it requires a set of texts with noun phrases annotated with the slot type to be extracted.

To adapt the AutoSlog approach to Korean, we first devised Korean equivalents of the English patterns, two of which are shown in Figure 3. It turned out that for our corpus, we could collapse some of these patterns, though some new ones were also needed. In the end we used just nine generic patterns.

Important issues that arose in adapting the approach were (1) greater flexibility in word order and heavier reliance on morphological cues in Korean, and (2) the predominance of light verbs (verbs with little semantic content of their own) and aspectual verbs in the chosen domain. We discuss these issues in the next two sections.

4.3 Korean Parser

We used Yoon's hybrid statistical Korean parser (Yoon et al., 1997, 1999; Yoon, 1999) to process the input sentences prior to extraction. The parser incorporates a POS tagger and morphological analyzer and yields a dependency representation as its output.[4] The use of a dependency representation enabled us to handle the greater flexibility in word order in Korean.

To facilitate pattern matching, we wrote a simple program to convert the parser's output to XML form. During the XML conversion, two simple heuristics were applied, one to recover implicit subjects, and another to correct a recurring misanalysis of noun compounds.

4.4 Trigger Word Filtering and Generalization

In the newswire corpus we looked at, meeting events were rarely described with the verb 'mannata' ('to meet'). Instead, they were usually described with a noun that stands for 'meeting' and a light or aspectual verb, for example, 'hoyuy-lul kacta' ('to have a meeting') or 'hoyuy-lul machita' ('to finish a meeting'). In order to acquire extraction patterns that made appropriate use of such collocations, we decided to go beyond the AutoSlog approach and explicitly group trigger words (such as 'hoyuy') into classes, and to likewise group any collocations, such as those involving light verbs or aspectual verbs. To find collocations for the trigger words, we reviewed a Korean lexical co-occurrence base which was constructed from a corpus of 40 million words (Yoon et al., 1997). We then used the resulting specification to filter the learned patterns to just those containing the

[3] For TIDES, we plan to use more sophisticated learning algorithms, as well as *active learning* techniques, such as those described in Thompson et al. (1999).

[4] Overall dependency precision is reported to be 89.4% (Yoon, 1999).

trigger words or trigger word collocations, as well as to generalize the patterns to the word class level. Because the number of trigger words is small, this specification can be done quickly, and soon pays off in terms of time saved in manually filtering the learned patterns.

4.5 Results

In testing our approach, we obtained overall results of 79% recall and 67% precision in a hold-one-out cross validation test. In a cross validation test, one repeatedly divides a corpus into different training and test sets, averaging the results; in the hold-one-out version, the system is tested on a held-out example after being trained on the rest. In the IE setting, the recall measure is the number of correct slots found divided by the total number of correct slots, while the precision measure is the number of correct slots found divided by the total number of slots found.

While direct comparisons with the MUC conference results cannot be made for the reasons we gave above, we nevertheless consider these results quite promising, as these scores exceed the best scores reported at MUC-6 on the scenario template task.[5]

Table 1: Hold-One-Out Cross Validation

Slots	Recall	Precision
All	79%	67%
Participant	75%	84%
Date/Location	86%	54%

Table 2: Hold-One-Out Cross Validation without Generalization

Slots	Recall	Precision
All	61%	64%
Participant	57%	81%
Date/Location	67%	52%

A breakdown by slot is shown in Table 1. We may note that precision is low for date and location slots because we used a simplistic sentence-level merge, rather than dependencies. To measure the impact of our approach to generalization, we may compare the results in Table 1 with those shown in Table 2, where generalization is not used. As can be seen, the generalization step adds substantially to overall recall.

To illustrate the effect of generalization, consider the pattern to extract the subject NP of the light verb 'kac (hold)' when paired with an object NP headed by the noun 'hyepsang (negotiation)'. Since this pattern only occurs once in our corpus, the slot is not successfully extracted in the cross-validation test without generalization. However, since this example does fall under the more generalized pattern of extracting the subject NP of a verb in the light verb class when paired with an object NP headed by a noun the 'hoytam-hyepsang' class, the slot is successfully extracted in the cross-validation test using the generalized patterns. Cases like these are the source of the 18% boost in recall of participant slots, from 57% to 75%.

5 Discussion

Our feasibility study has focused our attention on several questions concerning the interaction of IE and MT, which we hope to pursue under the DARPA TIDES initiative. One question is the extent to which slot filler translation is more practicable than general-purpose MT; one would expect to achieve much higher quality on slot fillers, as they are typically relatively brief noun phrases, and instantiation of a slot implies a degree of semantic classification. On the other hand, one might find that higher quality is required in order to take translated phrases out of their original context. Another question is how to automate the construction of bilingual lexicons. An important issue here will be how to combine information from different sources, given that automatically acquired lexical information is apt to be less reliable, though domain-specific.

Acknowledgements

Our thanks go to Richard Kittredge and Tanya Korelsky for helpful comments and advice. This work was supported by ARL contract DAAD17-99-C-0005.

5 http://www.nist.gov/itl/div894/894.02/related_projects/tipster/muc.htm

References

Cardie, C. (1997). Empirical Methods in Information Extraction. *AI Magazine* 18(4):65–79.

Lavoie, B. and Rambow, O. (1997). RealPro — A fast, portable sentence realizer. In *Proceedings of the Conference on Applied Natural Language Processing (ANLP'97)*, Washington, DC.

Lavoie, B., Korelsky, T., and Rambow, O. (2000). A Framework for MT and Multilingual NLG Systems Based on Uniform Lexico-Structural Processing. To appear in *Proceedings of the Sixth Conference on Applied Natural Language Processing (ANLP-2000)*, Seattle, WA.

Lehnert, W., Cardie, C., Fisher, D., McCarthy, J., Riloff, E., and Soderland, S. (1992). University of Massachusetts: Description of the CIRCUS system as used in MUC-4. In *Proceedings of the Fourth Message Understanding Conference (MUC-4)*, pages 282–288, San Mateo, CA. Morgan Kaufmann.

MUC-5 (1994). *Proceedings of the Fifth Message Understanding Conference (MUC-5).* Morgan Kaufmann, San Mateo, CA.

MUC-7 (1998). *Proceedings of the Seventh Message Understanding Conference (MUC-7).* Morgan Kaufmann, San Francisco, CA.

Nasr, A., Rambow, O., Palmer, M., and Rosenzweig, J. (1997). Enriching lexical transfer with cross-linguistic semantic features. In *Proceedings of the Interlingua Workshop at the MT Summit,* San Diego, CA.

Palmer, M., Rambow, O., and Nasr, A. (1998). Rapid prototyping of domain-specific machine translation systems. In *Machine Translation and the Information Soup - Proceedings of the Third Conference of the Association for Machine Translation in the Americas AMTA'98,* Springer Verlag (Lecture Notes in Artificial Intelligence No. 1529), Berlin.

Riloff, E. (1993). Automatically constructing a dictionary for information extraction tasks. In *Proceedings of the Eleventh National Conference on Artificial Intelligence,* pages 811–816, Washington, DC. AAAI Press / MIT Press.

Thompson, C. A., Califf, M. E., and Mooney, R. J. (1999). Active learning for natural language parsing and information extraction. In *Proceedings of the Sixteenth International Machine Learning Conference (ICML-99)*, Bled, Slovenia.

White, M. and Caldwell, T. (1998). EXEMPLARS: A practical, extensible framework for dynamic text generation. In *Proceedings of the 8th International Workshop on Natural Language Generation,* Niagara-on-the-Lake, Ontario.

Yoon, J. (1999). Efficient dependency parsing based on three types of chunking and lexical association. Submitted.

Yoon, J., Choi, K.-S., and Song, M. (1999). Three types of chunking in Korean and dependency analysis based on lexical association. In *Proceedings of ICCPOL.*

Yoon, J., Kim, S., and Song, M. (1997). New parsing method using global association table. In *Proceedings of the 5th International Workshop on Parsing Technology.*