# CUSTOMIZING AND EVALUATING A MULTILINGUAL DISCOURSE MODULE

**Chinatsu Aone**

System Research and Applications Corporation (SRA)
2000 15th Street North
Arlington, VA 22201
email: aonec@sra.com

## ABSTRACT

In this paper, we first describe how we have customized our data-driven multilingual discourse module within our text understanding system for different languages and for a particular NLP application by utilizing hierarchically organized discourse KB's. Then, we report quantitative and qualitative findings from evaluating the system both with and without discourse processing, and discuss how resolving certain kinds of anaphora affects system performance.

## 1 INTRODUCTION

Although previous discourse research (cf. Hobbs [7], Webber [9], Grosz and Sidner [6], etc.) made significant contributions at a theoretical level, the effectiveness of discourse processing in NLP systems has not been studied so far at a practical level (cf. Walker [8]). In systems used in NLP applications such as the Message Understanding Conferences (cf. [4, 5]), discourse processing is often not a separate module but is part and parcel of "template generation." Thus, the effect of different types of discourse processing on a particular task has not been shown either.

In addition, both at theoretical and practical levels, few seem to have considered designing discourse processing in a way that is customizable for multiple languages and domains. However, since discourse phenomena differ among languages and even among domains within the same language, it is desirable that discourse processing be customizable and its result evaluable.

In this paper, we describe how we have customized our multilingual discourse module within our text understanding system for a particular task (i.e. data extraction in the joint venture domain) in two different languages (i.e. English and Japanese), and report the evaluation results.

## 2 DISCOURSE MODULE ARCHITECTURE

In Aone and McKee [2], we have described our new language- and domain-independent discourse module within our text understanding system. In addition to being language- and domain-independent, the module is evaluable and trainable to different applications and domains. The discourse architecture is motivated by our need to port our text understanding system to different languages (e.g. English, Japanese, Spanish) and to different domains (cf. Aone et al. [1]). The discourse module is strictly data-driven so that anaphora resolution for different languages and domains can be achieved simply by selecting necessary data. It consists of one discourse processor (the Resolution Engine) and three discourse knowledge bases (the Discourse Phenomenon KB, the Discourse Knowledge Source KB, the Discourse Domain KB). The Discourse Administrator is a development-time tool for defining the three discourse KB's. The architecture is shown in Figure 1.

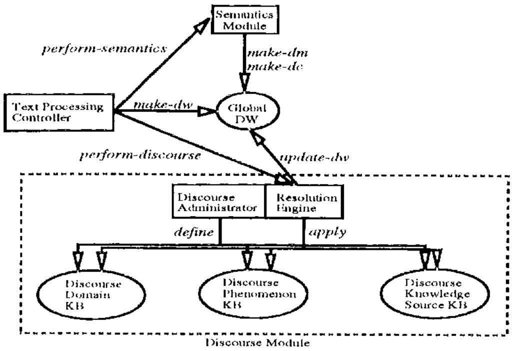

Figure 1. Discourse Architecture

### 2.1 Discourse Knowledge Bases

The Discourse Knowledge Source KB houses small well-defined anaphora resolution strategies. Each knowledge source (KS) is an object in the hierarchically organized KB, and information can be inherited from more general to more specific KS's. This KB consists of three kinds of KS's: generators, filters and orderers. A *generator* is used to generate possible antecedent hypotheses from a certain region of text. A *filter* is used to eliminate impossible hypotheses, while an *orderer* is used to rank possible hypotheses in a preference order if there is more than one.

Most of the KS's are language-independent (e.g. all the generators and the semantic filters). Even when they are language-specific, a sub-KS can inherit information from its superclass KS's while defining specific data locally. For example, the Semantic-Gender-Filter KS[1] defines only functional definition of this KS, while its sub-KS's for English and Japanese each specify language-specific data and inherit the same functional definition from their parent KS.

[1] Semantic-Gender-Filter filters out an antecedent hypothesis whose semantic gender is not consistent with the restriction imposed by the syntactic gender of a pronoun.

The Discourse Phenomenon KB contains hierarchically organized discourse phenomenon objects (e.g. Name-Anaphora, Definite-NP) each of which specifies a definition of the discourse phenomenon and a set of KS's (i.e. generators, filters, and orderers) to apply to resolve this particular discourse phenomenon. Because the discourse KS's are independent of discourse phenomena, the same discourse KS can be shared by different discourse phenomena in different languages and domains. For example, KS's such as Semantic-Type-Filter and Recency-Orderer are used by most discourse phenomena in multiple languages.

Finally, the Discourse Domain KB contains discourse domain objects each of which defines a set of discourse phenomena to handle in a particular domain. Since texts in different domains exhibit different sets of discourse phenomena, and since different applications even within the same domain may not have to handle the same set of discourse phenomena, the discourse domain KB is a way to *customize* and *constrain* the workload of the discourse module.

These three hierarchically organized discourse KB's make it possible to share some of the discourse KB's while also being able to add language- and domain-specific discourse data.

### 2.2 Resolution Engine

The Resolution Engine is the run-time processing module which finds the best antecedent hypothesis for a given anaphor by using the discourse KB's described above. First, it determines from the Discourse Domain KB which discourse phenomena to handle given a particular language and domain. Then, it uses the Discourse Phenomenon KB to classify an anaphor as one of the discourse phenomena and to decide which KS's to apply to it. Next, the Engine applies appropriate generator KS's to get an initial set of antecedent hypotheses, and then applies filter KS's to remove inconsistent hypotheses. When there is more than one hypothesis left, orderer KS's specified in the Discourse Phenomenon KB are invoked to rank the hypotheses.

## 3 CUSTOMIZING DISCOURSE KB'S

We have customized our discourse KB's to perform a data extraction task in the joint venture domain. Our text understanding system takes English and Japanese newspaper articles about joint ventures as input (cf. Figure 2), and outputs database templates (cf. Figure 3). The system has to extract from the articles information regarding which organizations participate in a joint venture (including a new joint venture company if any), what the purpose of the joint venture is (e.g. selling coal), who the people are that are associated with these organizations, etc. We made a task-oriented decision that handling *organization* anaphora, both *definite NPs* (e.g. "the company") and *name anaphora* (e.g. "Toyota" for "Toyota Motors Corp."), is a top priority initially in order to improve performance. Thus, we created in the Discourse Domain KB a discourse domain object called JV-Data-Extraction which specifies that two discourse phenomenon objects from the Discourse Phenomenon KB, namely name anaphora (DP-Name) and definite NP anaphora for organizations (DP-DNP-Organization), should be handled in this application domain.

> NEW YORK -- A joint venture to export coal from the United States has been formed between M&M Ferrous America Ltd. here and Crown Coal & Coke Co., Pittsburgh.
>
> Coal obtained by Crown from various domestic mines will be marketed offshore by M&M, a trading company formed six years ago by former Philipp Brothers Inc. employees. Crown, which formerly had its own mines, heretofore marketed coal from various sources to domestic steelmakers only, according to Eric S. Katzenstein, M&M vice president.
>
> ((omitted))
>
> Eastern European countries such as Romania are likely markets, he said.

Figure 2. An Example of Input Text

```
<TIE_UP_RELATIONSHIP-2975348-1> :=
  TIE-UP STATUS: EXISTING
   ENTITY: <ENTITY-2975348-1> <ENTITY-2975348-2>
   ACTIVITY: <ACTIVITY-2975348-1>

<ENTITY-2975348-1> :=
  NAME: M&M Ferrous America LTD
  ALIASES: "M&M"
  LOCATION: New York (CITY 4) New York (PROVINCE 1)
  United States (COUNTRY)
  TYPE: COMPANY
   PERSON: <PERSON-2975348-1>

<ENTITY-2975348-2> :=
  NAME: Crown Coal & Coke CO
  ALIASES: "Crown"
  LOCATION: Pittsburgh (CITY 4) Pennsylvania (PROVINCE 1)
  United States (COUNTRY)
  TYPE: COMPANY

<INDUSTRY-2975348-1> :=
  INDUSTRY-TYPE: SALES
  PRODUCT/SERVICE: (50 "Crown's coal")

<ACTIVITY-2975348-1> :=
   INDUSTRY: <INDUSTRY-2975348-1>
  ACTIVITY-SITE: (Romania (COUNTRY) <ENTITY-2975348-
  1>)

<PERSON-2975348-1> :=
   NAME: Eric S. Katzenstein
  PERSON'S ENTITY: <ENTITY-2975348-1>
  POSITION: SREXEC
```

Figure 3. An Example of an Output Template

### 3.1 Name Anaphora

In order to resolve name anaphora, English and Japanese *share* some of the KS's in the Discourse Knowledge Source KB, namely Current-Text-Generator, Semantic-

Type-Filter, and Recency-Orderer. This generator generates all the possible antecedent hypotheses up to the current sentence. The Semantic-Type-Filter then checks if the semantic type of anaphor is consistent with that of an antecedent hypothesis. When there is more than one hypothesis left, the Recency-Orderer orders the hypotheses according to their proximity to the anaphor.

In addition to the three language-independent KS's, each language uses a language-specific filter. For English, a filter named English-Name-Filter, which matches an anaphor (e.g. "Crown") with a subsequence of an antecedent name string (e.g. "Crown Coal & Coke CO"), is currently employed. For Japanese, an additional single filter called Japanese-Name-Filter covers seemingly vast variations of Japanese company name anaphora[2]. This KS matches an anaphor with any combination of *characters* in an antecedent as long as the character order is preserved (e.g. "abc" can be an anaphor of "abcde"). One exception is that an anaphor can have an extra word "sha" at the end that is not a part of the full company name or a company acronym (e.g. "Westinghouse (WH)" can be referred to anaphorically by "Westinghouse-sha" or "WH-sha").

### 3.2 Definite NP

Another discourse phenomenon which is handled for this task is definite NPs referring to *organizations* such as "the venture," "the West German electronics concern," etc., where the words "venture" and "concern" in these contexts point to subclasses of the semantic concept for an organization. Although Japanese does not have a definite article, in written Japanese the word "dou" (literally meaning "the same") prefixed to certain nouns performs approximately the same function as English definite article "the". Both English and Japanese currently *share* the same three KS's (i.e. Current-Text-Generator, Semantic-Type-Filter, Recency-Orderer) for definite NP resolution.

Additionally, English uses Syntactic-Number-Filter, which checks if the syntactic number of the anaphor is consistent with that of an antecedent hypothesis. Although Japanese does not exhibit syntactic number distinction, a "dou" phrase can only refer semantically to a single entity.[3] Thus, Japanese uses Semantic-Amount-Filter, which excludes semantically plural entities (e.g. a conjoined NP, an NP with a plural quantifier) as possible antecedents for a "dou" phrase.

## 4 EVALUATION RESULTS

In this section, we will report our evaluation results. We ran 100 Japanese and 100 English blind test joint venture articles through our text understanding system with and without the discourse module turned on, and scored the results using an automatic scoring program. The scoring program uses a scoring metric from information retrieval, and reports *recall* and *precision* for each slot in the templates as well as a single combined score called *F-measure*[4] for overall performance (cf. [4]).

It should be noted that this evaluation is a *blackbox* evaluation of the system as used in a particular application task. Consequently, the results do not directly reflect the performance of the discourse module itself. For example, this task does not require all company name anaphora (i.e. aliases) to be reported, but only those which are involved in joint ventures. Also, the causes of task failure or success are sometimes due to the failure or success of system modules other than the discourse module. For instance, the preprocessing system does not always recognize company names which are potential antecedents. On the other hand, the preprocessing module rather than the discourse module sometimes recognizes company acronyms as aliases. Thus, the results of the blackbox evaluation reflect more on how the discourse module helps the whole system perform a particular task.

### 4.1 Name Anaphora

It is clear that the performance of name anaphora resolution is directly linked to how well the system fills in the ALIASES slot in the output templates (cf. Figure 3). The 100 Japanese texts required identifying a total of 127 company name aliases. With the discourse module turned on, the recall of the ALIASES slot increases by 38 points and the precision by 16 points. Though the set of KS's used for name anaphora was mostly satisfactory, we found one problem particular to this domain in both languages. Since the texts are in the joint venture domain, it is often the case that the name of a new joint venture company (e.g. "Chrysler Japan") overlaps the names of its parent companies (e.g. "Chrysler Corp."). When the text uses a name anaphor (e.g. "Chrysler"), it must refer to the parent company even when the joint venture company is mentioned most recently. We are planning to add another orderer which prefers the parent company when there is such a conflict.

### 4.2 Definite NP

We hypothesized that resolving definite NP's affects the extraction of information about which company is performing which "economic activity" in a joint venture (e.g. Company A will manufacture cars while Company B will market them), since such information appears later in an

[2] For example:

全日空(全日本空輸), 東京相銀(東京相互銀行),
ロ社(ロッキード社), ネイチャー社(ザ・ネイチャーカンパニー).

[3] A definite plural NP can be expressed in Japanese by a numeral or numerical quantifier plus a classifier, as in "ryousha" (the two companies) and "san-sha" (the three companies).

[4] F-measure is calculated by:

$$F = \frac{(\beta^2 + 1.0) \times P \times R}{\beta^2 \times P + R}$$

where $P$ is precision, $R$ is recall, and $\beta$ is the relative importance given to recall over precision. In this case, $\beta = 1.0$.

article after companies involved in a joint venture are already introduced into the discourse (e.g. "Publishing rivals Time Inc. and New York Times Co. said they agreed in principle to form a jointly owned national magazine distribution partnership... *The joint venture* will continue to market magazines currently marketed by Time Distribution...").

Under the same test condition as above, the precision of the relevant slot (i.e. ACTIVITY-SITE slot in Figure 3) increased by 5 points in Japanese when discourse processing was used. The recall was not affected much by the discourse processing; it increased only by 1 point. In the English test, the changes in both precision and recall were negligible. One of the reasons for this less drastic increase of this slot value is that the sentence expressing economic activities do not always use definite NPs for the agents of such activities. Such agents can be expressed by name anaphora or pronouns or, often in English, by implicit subjects of infinitives, as in "Siemens AG and GTE Corp. agreed to set up a new holding company in West Germany *to oversee their telecommunications joint venture...*".

In addition, examination of the test results showed that when there are more than one antecedent hypothesis, topic marking (using particle "wa") plays a more significant role in determining the antecedent of a Japanese "dou" definite NP than recency. At the time of the testing, however, we were not using topic marking information to prefer topicalized antecedent hypotheses. Another finding which is true of both Japanese and English is that definite NP anaphora resolution often requires pragmatic inferencing in order to obtain a fact which is not explicitly stated in the text. For example, in order to resolve the definite NP in the sentence "Chevron, an oil company, also said it acquired Rhone-Poulenc's 30% interest in Petrosynthese S.A., boosting its holding in *the French joint venture* to 65%," the discourse module has to infer either that Petrosynthese S.A. is a French company (perhaps from the company designator?) or that acquiring someone's holding in a company increases one's holding in that company. We are currently adding KS's which make use of topic information and pragmatic inferencing, and also investigating which combinations of KS's will optimize discourse performance.

Furthermore, we think that very little change in recall is due to the fact that the system assumed the parent companies to be the value of ACTIVITY-SITE when it is undetermined. Thus, this default value kept the recall of the system without discourse processing higher, and therefore the ACTIVITY-SITE slot was not as good an indicator of the discourse module performance as the ALIASES slot.

It is interesting to note that an approach like Dagan and Itai's [3], which uses statistical data on semantic selectional restriction that is automatically acquired from large corpora to resolve anaphora[5], does not work well in this domain. This is because a typical text in this domain contains at least two possible antecedents (joint venture partners and possibly a joint venture company) of the same semantic type, namely organization, for a definite NP anaphora referring to organizations.

### 4.3 Overall Performance

Overall, discourse processing increased the system performance measured by the combination of overall recall and precision scores (i.e. F-measure) by 4 points in Japanese, mostly due to an overall increase in *precision.* Interestingly, the discourse processing helped also in the identification of *links* between organizations and people, as indicated by the PERSON slot of the <ENTITY> object and the PERSON'S ENTITY slot of the <PERSON> object (cf. Figure 3). With the discourse processing turned on, the recall of both PERSON and PERSON'S ENTITY slots increased by 7 points, and the precision by 10 points and 12 points respectively.

We think that this is because when a person associated with an organization is mentioned, the company name or the person's name is often an anaphoric form as in "Carlos M. Herrera, president of *Preferred,*" or "*Katzenstein,* a former executive with Bomar Resources Inc.". In order to understand the relation between an organization and a person as in "Eric S. Katzenstein, M&M vice president" (cf. Figure 2), the system has to recognize both the affiliation link between the person and the company implicit in the appositive phrase, and the anaphoric link between the objects under different aliases. Our discourse module takes care of both identifying appositive relations (e.g. Eric S. Katzenstein *is* vice president) and resolving name anaphora (e.g. "M&M" refers to "M&M Ferrous America Ltd.").

## 5 CURRENT AND FUTURE WORK

In this paper, we have described our multilingual discourse module and its customized discourse KB's, and reported the blackbox evaluation results when it was used in a data extraction task in the joint venture domain. Currently we are working on the following two research areas in order to improve anaphora resolution.

First, we are experimenting with ways to automate training of anaphora resolution by applying machine learning so that the discourse module can be customized automatically to a particular language, domain or application without extensive manual knowledge engineering. In order to obtain feedback for training, we must be able to automate *glassbox* evaluation of discourse processing itself. For this, we have built two tools: a discourse tagging tool and a discourse evaluation tool. The former has been used to tag texts with discourse relations, while the latter takes discourse-tagged corpora as a key and the system output as results to be evaluated.

[5] According to their approach, for a sentence "It was going to collect it," "government" is a preferred antecedent of the first "it," while "money" is of the second, using such statistics.

Second, we are expanding the range of anaphoric phenomena which our discourse module can handle. They include overt pronouns in English and Spanish, and zero pronouns in Japanese and Spanish.

## REFERENCES

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

[2] Chinatsu Aone and Douglas McKee. Language-Independent Anaphora Resolution System for Understanding Multilingual Texts. In *Proceedings of 31st Annual Meeting of the ACL*, 1993.

[3] Ido Dagan and Alon Itai. Automatic Acquisition of Constraints for the Resolution of Anaphora References and Syntactic Ambiguities. In *Proceedings of the 13th International Conference on Computational Linguistics*, 1990.

[4] Defense Advanced Research Projects Agency. *Proceedings of Fourth Message Understanding Conference (MUC-4)*. Morgan Kaufmann Publishers, 1992.

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

[6] Barbara Grosz and Candace L. Sidner. Attentions, Intentions and the Structure of Discourse. *Computational Linguistics*, 12, 1986.

[7] Jerry R. Hobbs. Pronoun Resolution. Technical Report 76-1, Department of Computer Science, City College, City University of New York, 1976.

[8] Marilyn A. Walker. Evaluating Discourse Processing Algorithms. In *Proceedings of 27th Annual Meeting of the ACL*, 1989.

[9] Bonnie Webber. A Formal Approach to Discourse Anaphora. Technical report, Bolt, Beranek, and Newman, 1978.