Deriving Semantic Knowledge from Descriptive Texts using an MT System

Eric Nyberg[1], Teruko Mitamura[1], Kathryn Baker[1],
David Svoboda[1], Brian Peterson[2], and Jennifer Williams[2]

[1] Language Technologies Institute
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
{ehn,teruko,klb,svoboda}@cs.cmu.edu

[2] Ontology Works, Inc.
1132 Annapolis Rd, Suite 104
Odenton, MD 21113-1672
{peterson,williams}@ontologyworks.com

Abstract. This paper describes the results of a feasibility study which focused on deriving semantic networks from descriptive texts using controlled language. The KANT system [3,6] was used to analyze input paragraphs, producing sentence-level interlingua representations. The interlinguas were merged to construct a paragraph-level representation, which was used to create a semantic network in Conceptual Graph (CG) [1] format. The interlinguas are also translated (using the KANTOO generator) into OWL statements for entry into the Ontology Works electrical power factbase [9]. The system was extended to allow simple querying in natural language.

1 Introduction

This paper reports on a study which adapted machine translation tools to generate knowledge representation languages as part of a knowledge extraction system. The inputs to the system are short texts describing critical infrastructures (e.g. financial markets, electrical power transmission). Where necessary, the texts are rewritten to conform to a controlled language [4]. Then each text is analyzed to produce an interlingua representation (IR) for its sentences. The interlinguas are merged together into a single representation for the paragraph. The merged representation is then generated into different output languages. In this study, the outputs were not generated in natural language, but as statements in different knowledge representation languages - Conceptual Graphs (CG) / Knowledge Interchange Format (KIF) [1], and OWL [9] (see Figure 1).

We used the KANTOO MT system [7] for the analysis and generation steps. We investigated two output representations: Conceptual Graph / Knowledge Interchange Format (KIF) [1], and the Ontology Works OWL language [9]. The study focused on an ontology and textual description for a model of the Northwest electric power grid [10]. A set of texts were written to describe the elements of the model and their attributes; these texts were re-written to conform to a controlled language. The KANTOO analyzer was extended to produce merged interlingua structures for these texts, and the KANTOO generator was extended with special knowledge for generating from merged interlinguas into KIF and OWL. The OWL statements were instantiated as facts in an Ontology Works factbase, instantiating concepts in an upper model for the electric power domain.

In an extension of the basic system, we enhanced the KANTOO analyzer with the ability to process natural language questions about the concepts in the factbase. After translating these questions into the appropriate OWL queries, KANTOO was able to query the factbase and return the appropriate results.

In Section 2 we describe the controlled language analysis used to create interlinguas for the input sentences. Section 3 outlines the merging algorithm used to create semantic graphs from the interlinguas. In Section 4 we discuss the generation of KIF and OWL output from the interlingua. Section 5 discusses an initial capability for natural language querying of the resulting fact base. We conclude in Section 6 with some remarks about open issues and possible future work.

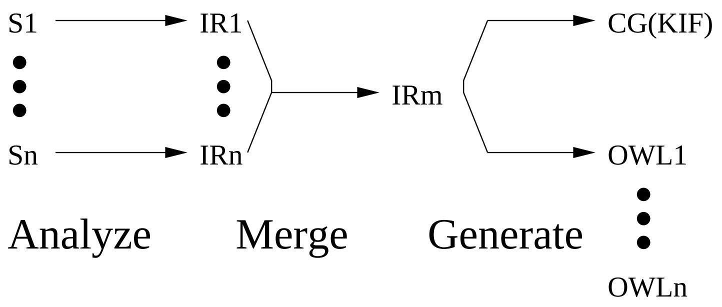

Fig. 1. Three-Step Processing

2 Controlled Language Analysis

The general goals of Controlled Language (CL) are to achieve consistent authoring of source texts and to encourage clear and direct writing. Controlled language is also used to improve the quality of translation output through the use of short, concise and unambiguous sentences [4, 2, 8]. Although the goal of this study was to generate knowledge representations, rather than translated texts, this task also requires an accurate, unambiguous meaning representation.

2.1 Controlled Language in KANT

In KANT Controlled English (KCE) we explicitly encode domain lexemes along with their meanings. Whenever possible, the lexicon should encode a single meaning for each word/part-of-speech pair. When a term must carry more than one meaning in the domain, these meanings must be encoded in a separate lexical entry. If more than one such entry is activated during analysis, and the ambiguity cannot be resolved by the grammar, then interactive disambiguation is used to select the intended meaning. The analysis grammar encodes a set of writing guidelines and constraints which reduce ambiguity, both at the phrasal level and at the sentential level (for more details, see [4]).

2.2 The KANT Interlingua

The KANT analyzer uses a lexicon and a unification grammar to produce a set of LFG-style f-structures. These f-structures are mapping into interlingua expressions by a semantic interpreter module. Each interlingua may contain semantic concepts, semantic features, and semantic roles representing the meaning of various lexical and structural components of the input sentence. Some grammatical information may also be preserved in the interlingua, if it is necessary for accurate translation. An example interlingua is shown in Figure 2.

```
(*A-DETERMINE
  (argument-class agent+patient)
  (mood imperative)
  (patient
    (*O-VULNERABILITY
      (distance near)
      (number mass)
      (q-modifier
        (*Q-characteristic_WITH
          (case
            (*K-WITH))
          (object
            (*O-PB-TRADER
              (number plural)
              (reference no-reference)))
          (role characteristic)))
      (reference definite)))
  (punctuation period)
  (tense present))
```

Fig. 2. KANT Interlingua for *Determine this vulnerability with PB traders*

In KANT, the concept is lexically specified by the head of the syntactic constituent that the interlingua corresponds to. A semantic role is a slot that is filled with an embedded interlingua expression. The embedded interlingua is headed by the concept associated with the head of the syntactic constituent that the semantic role corresponds to. For example, `subject` and `object` in the f-structure can correspond to `agent` and `theme` in the interlingua.

2.3 Designing a Controlled Language for Descriptive Texts

One primary objective of the study was to design a CL grammar whose interlingua output could be mapped into knowledge representation languages. The first step was to determine a set of paragraph-level writing guidelines, by rewriting sample texts into a form suitable for automatic analysis. The rewriting strategies we used can be thought of as an initial specification for CL authoring at the paragraph level. This is a significant departure from past work in CL, where the unit of analysis is typically a sentence. The sample texts were drawn from two different domains: economic system vulnerabilty and power system vulnerability (see Figure 3). The first step in applying a CL to a new domain is to add unfamiliar terms to the lexicon. We used our existing English lexicon and added new terms from the economic and power domains.

Original:
Vulnerability of the economic system Sigma sub H2 was studied by determining the U-vulnerability and Lambda-vulnerability of a few simple system instantiations.
Rewritten:
Determine U-vulnerability and Lambda-vulnerability of a few simple instantiations of Sigma sub H2, which is an economic system.

Original:
The U-vulnerability for an ST exchange rate model instantiation of Sigma sub H2 was examined for the case in which the currency trader estimate of "true" exchange rate is manipulated. For this instantiation, and with "output" defined to be GDP volatility magnitude, U-vulnerability was determined using signal bounding arguments.
Rewritten:
The U-vulnerability of the ST exchange rate model Sigma sub H2 was determined with the output as GDP volatility magnitude. This U-vulnerability was determined by manipulation of the trader estimate of the true exchange rate. This U-vulnerability was determined by using signal bounding arguments.

Fig. 3. Example Texts Rewritten to CL

We experimented with two different approaches to rewriting the source texts. First, we took the same approach that we have used for past domains, where an existing text is rewritten on a sentence by sentence basis, according to the KANT CL Specification. This method focuses on grammatical constructions. We found that while this approach was not particularly difficult, the interlingua output did not necessarily model the semantic relations required to build the graph structures required for the CG output. To reduce the complexity of mapping interlingua into CG format, it is necessary to promote isomorphism, in the sense that meanings which are represented with identical structures in CG should be represented using parallel structures in the KCE input and interlingua.

The second approach to rewriting was to define the simplest canonical sentence structure for each graph relation produced in the first approach, working backward from graph to text. The original English sentences were used only for clarification or verification of meaning. This novel method of working from graph to text is central to the success of the study. This method has the advantage of addressing several sentences at once, because a large graph conveys the meaning of more than one sentence at a time. Graphs need not be tied to a sentence-level analysis. Another advantage is that we were able to produce simple but meaningful sentences, even though our researchers were unfamiliar with the domain. Non-domain experts should find the process accessible.

For these reasons, we decided to use this graph-based approach to rewriting for the rest of the study. We continued to use sample graphs as a primary reference in designing the CL sentence structures.

2.4 Refining the CL for Paragraphs

We began to look at shared elements from sentence to sentence, in order to reason about how the interlinguas from two or more sentences might be merged. We established some initial guidelines for structuring the text around CG creation. In the field of CL generally, and also for graph-based CL, it is essential to follow the guidelines consistently in order to obtain the best possible output.

One guideline is to repeat the same verb and object from a single graph with different modifiers. For example, if the graph is headed by the action *Determine* and the theme is *Lambda-vulnerability*, then the phrase *Determine Lambda-vulnerability* can be repeated with the `Chrc` (characteristic) *MM & MA traders* and with the `Rslt` (result) *output as GDP volatility magnitude.*

- **(Before)** Determine this Lambda vulnerability *with MM traders and MA traders.*
- **(After)** Determine this Lambda vulnerability *with output as GDP volatility magnitude.*

Another writing guideline requires the writer to refer to a mass noun, such as *U-vulnerability*, by using a determiner such as *the* or *this*, once the noun has been introduced into the context:

- **(Before)** U-vulnerability of the ST exchange rate model Sigma sub H2 was determined with the output as GDP volatility magnitude.
- **(After)** *This U-vulnerability* was determined by manipulation of the trader estimate of the true exchange rate.

Other ongoing work includes developing a set of *key words*. Key words written in the text can correspond to graph node creation. For example, *in order to* signals the graph node `Purpose`. The phrase *By using* signals the graph node `Method`. Use of key words is important because it enables directed writing by the authors and also makes the graphs more intuitive and clear.

Once a set of IRs is produced, the goal is to combine them into a single IR structure that corresponds to a unique graph. Features in an IR which were superfluous for this goal, such as syntactic features which would be used by a translation system, were identified as features which could be omitted. As patterns for merging were identified, these were factored into the merging algorithm, which will be described in the following section.

3 Merging Interlinguas into CG

The interlingua merging algorithm produces a single conceptual graph (CG) output for the set of interlingus (IRs) produced by KANT for a particular text. In this section, we describe the steps that are taken to produce the merged IRs:

1. **Preprocess the IRs**. This entails removing some structural slots that exist in the Interlingua Representations (IRs) that aren't necessary in order to construct CG output. For example, certain grouping structures such as `*G-COORDINATION` are replaced with an ordered list of slot fillers.
2. **Construct an 'IR Index' for Every Concept**. Indexing each concept is an important step towards finding similar IRs to merge into one in the later step. The means of finding the concept is represented using a path of slots. We can use this index to locate and compare a set of sub-IRs that all start with the same concept.
3. **Find Concepts That Appear In More Than One IR**. To merge the IRs, we must identify IRs which can be unified; i.e., their meaning are consistent and can be combined into a single graph. A concept has multiple instances if it heads two or more sub-IRs that cannot be unified. These sub-IRs can be determined using the IR Index generated in Step 2. If all the sub-IRs for a concept unify, the concept has only one unique instance. Otherwise, the number of instances of the concept is equal to the minimum number of sub-interlinguas that cannot be unified.
   This simple type of unification (via concept equivalence) might be extended to unification under subsumption, or by merging concepts using coordination; e.g., the concepts *O- TRADER- ESTIMATE and *O- TRADER-EXCHANGE- RATE- ESTIMATE can be unified rather than treated as separate instances of different concepts.
4. **Create Unique Identifiers for Multiple Instances**. Any remaining instances of the same concept (e.g. `*O-MANIPULATION`) are re-labelled with unique indices, e.g. `*O-MANIPULATION-a` and `*O-MANIPULATION-b`.
5. **Create a Unified IR For Each Concept**. At this point, we can guarantee that each concept identifier used in the IRs has exactly one instance. We associate with each concept identifier an IR that is the unification of all sub-IRs that modify that concept. Since we store each unified IR separately, we can replace any occurrence of an IR as slot filler in another IR with a simple reference (pointer).
6. **Merge The Unified IRs**. At this point we traverse the unified IRs, and substitute other unified IRs for lone concept indices (pointers).
7. **Remove Slots That Don't Contain Concepts**. Remaining features which don't contain concepts (e.g., number, reference) are removed.

The merging algorithm returns all the IRs generated by following these steps in sequence. In the next section, we discuss how the merged IR is generated into KIF and OWL.

4 Generating Knowledge Representations: KIF and OWL

The structures created by the IR merging process are generated as KIF and OWL. Since the two representations are different, we address each in turn. The merged interlingua expressions can be directly generated into CG/KIF form by associating a node with each concept, and an arc with each semantic relation. The merged interlingua expressions are isomorphic to KIF structures, with the exception of certain tranformations which are required to map multiply-filled slots in the IR to multiple edges in the KIF graph. The KIF file format utilized also requires that each KIF graph be output on a single line. Due to the inherent near-isomorphism, it was relatively simple to generate KIF output from the merged IR. The generation of KIF structures from IR is completely automatic; the semantic relations in the KIF graph are taken directly from the semantic roles and features in the IR. An example of a completed KIF output can be seen in Figure 4.

```
(*PROP-RAVER
 (comparison-theme
  (*O-COMBUSTION-POWER-PLANT
   (located_IN
    (*PROP-PACIFIC-NORTHWEST-POWER-GRID
     (comparison-theme
      (*O-POWER-GRID
       (attribute (*P-ELECTRICAL)))))))
 (possesses
  (*O-GENERATION-OUTPUT-e
   (attribute (*P-MAXIMAL))
   (value_OF
    (*U-MEGAWATT-e (quantity (``2200''))))))
```

Fig. 4. Example KIF Graph (ASCII format)

The Ontology Works fact base is implemented as a relational database, with an API that allows semantic predicates and relations to be instantiated via a knowledge representation language called OWL [9]. In order to generate merged IRs as OWL statements, each IR was decomposed into the appropriate set of semantic primitives. The merged interlinguas for the entire text were generated into OWL form, using the KANTOO generator module. These OWL statements are then passed to the Ontology Works batch loader for insertion into the fact base. An example OWL output for the IR in Figure 4 is shown in Figure 5.

```
(EPctx.EPGrid PNWgrid)
(EPctx.CombustionPowerPlant RAVER)
(EPctx.epPartOf RAVER PNWgrid)
(EPctx.maxRatedGenerationOutput RAVER (MetricCtx.megawatt 2200))
```

Fig. 5. Example OWL Output

5 Querying the Fact Base in Natural Language

Once the semantic information has been extracted from the source text and loaded into the Ontology Works fact base, various queries can be run against the fact base to examine the information that was extracted. OWL includes a query language which can be used to formulate database queries. The initial version of our system can also accept natural language queries, which are mapped to interlingua form by the KANTOO analyzer, and generated as OWL queries by the KANTOO generator. A sample text is shown in Figure 6, and some example queries are shown in Figure 7.

```
The Pacific Northwest power grid (PNW power grid) is an electrical power
grid. Custer, Monroe, Paul, Allston and Keeler are thermal power plants
in the PNW power grid. Custer, Monroe, Paul, Allston and Keeler have a
maximal generation output of 650 megawatts.

Raver is a combustion power plant in the PNW power grid. The Dalles is
a hydro power plant in the PNW power grid. Raver has a maximal generation
output of 2200 megawatts. The Dalles has a maximal generation output of
1807 megawatts.

Grand Coulee and Chief Joseph are power plants in the PNW power grid.
Grand Coulee and Chief Joseph are members of the Upper Columbia power
plants group. The Upper Columbia power plants group is part of the PNW
power grid. Grand Coulee has a maximal generation output of 6480
megawatts. Chief Joseph has a maximal generation output of 2520 megawatts.

The Lower Columbia power plants group is part of the PNW power grid. John
Day and The Dalles are part of the Lower Columbia power plants group. John
Day has a maximal generation output of 1160 megawatts.
```

Fig. 6. Input Text: Northwest Power Grid Plants

```
What are the subparts of the PNW power grid?

(EPctx.epPartOf ?x PNWgrid)

Results:

  MALIN-ROUND-MOUNTAIN-POWER-LINE-a
  BIG-EDDY-THE-DALLES-POWER-LINE-a
  CUSTER-MONROE-POWER-LINE-a
  PAUL
  RAVER
  MALIN-ROUND-MOUNTAIN-POWER-LINE-b
  ALLSTON
  UPPER-COLUMBIA-POWER-PLANTS-GROUP
  CHIEF-JOSEPH
  CUSTER
  THE-DALLES
  GRAND-COULEE
  KEELER
  BIG-EDDY-THE-DALLES-POWER-LINE-b
  LOWER-COLUMBIA-POWER-PLANTS-GROUP
  MONROE
  CUSTER-MONROE-POWER-LINE-b

Are there any hydro power plants?

(EPctx.HydroPowerPlant ?x)

Results:

  JOHN-DAY
  THE-DALLES

The Dalles is connected to what?

(EPctx.epConnectsTo THE-DALLES ?x)

Results:

  BIG-EDDY-THE-DALLES-POWER-LINE-b
  BIG-EDDY-THE-DALLES-POWER-LINE-a
```

Fig. 7. Sample NL Queries, OWL and System Output

6 Conclusion

In this paper we have shown how an existing machine translation system can be adapted to generate output in a knowledge representation language instead of a human language. By combining this capability with a controlled language definition for specific domains (e.g., economic systems, power grids), it is feasible to create knowledge acquisition systems to populate a fact base from natural language texts. The KANTOO machine translation system was used to generate both KIF output structures and OWL statements to represent the facts in texts about economic systems and power grids.

The texts we have examined initially are simple descriptions of the static structures and relationships in the two domains. The system could be used to build and extend the ontology currently used by KANTOO for source text analysis. Most of the semantic knowledge used by the Analyzer is in the form of slot-filler restrictions [5], which could be learned by extracting relevant KIF fragments from texts as they are written for a new domain.

Future work should also focus on dynamic (non-monotonic) descriptions which require more reasoning (i.e., truth maintenance, conflict resolution) in the fact base as what is stated about the domain changes over time.

References

1. Hayes, P. and C. Menzel: A Semantics for the Knowledge Interchange Format, IJCAI 2001 Workshop on the IEEE Standard Upper Ontology, Aug. 6. (2001)
2. Kamprath, C., Adolphson, E., Mitamura, T. and Nyberg, E.: Controlled Language for Multilingual Document Production: Experience with Caterpillar Technical English. In: Proceedings of the Second International Workshop on Controlled Language Applications (1998)
3. Mitamura, T., Nyberg, E. and Carbonell, J.: An Efficient Interlingua Translation System for Multi-lingual Document Production. In: Proceedings of the Third Machine Translation Summit (1991)
4. Mitamura, T. and Nyberg, E.: Controlled English for Knowledge-Based MT: Experience with the KANT System. In: Proceedings of TMI-95 (1995)
5. Mitamura, T., Nyberg, E., Torrejon E. and Igo, R.: Multiple Strategies for Automatic Disambiguation in Technical Translation, Proceedings of TMI-99 (1999).
6. Nyberg, E. and Mitamura, T.: The KANT System: Fast, Accurate, High-Quality Translation in Practical Domains. In: Proceedings of COLING-92 (1992)
7. Nyberg, E. and T. Mitamura: The KANTOO Machine Translation Environment. In Proceedings of AMTA-2000.
8. Nyberg, E., T. Mitamura and W. Huijsen (to appear). "Controlled Language," in H. Somers, ed., *Computers and Translation: Handbook for Translators*, to be published by Johns Benjamins.
9. OWL and the IODE: The Ontology Works White Paper. Available at http://www.ontologyworks.com/whitepaper.pdf. August (2001).
10. Kosterev, D. N., C. W. Taylor and W. A. Mittelstadt: Model Validation for the August 10th, 1996 WSCC System Outage, IEEE Transactions on Power Systems, Volume 14, Number 03, August (1999).