The KANTOO Machine Translation Environment

Eric Nyberg and Teruko Mitamura

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

Abstract. In this paper we describe the KANTOO machine translation environment, a set of software services and tools for multilingual document production. KANTOO includes modules for source language analysis, target language generation, source terminology management, target terminology management, and knowledge source development. The KANTOO system represents a complete re-design and re-implementation of the KANT machine translation system.

1 Introduction

KANTOO is a knowledge-based, interlingual machine translation system for multilingual document production. KANTOO includes: a) an MT engine, the result of fundamental redesign and reimplementation of the core algorithms of the KANT system [2, 4]; and b) a set of off-line tools that support the creation and update of terminology and other knowledge resources for different MT applications. Several workflows are supported by KANTOO (see Figure 1):

- **Controlled-language authoring and checking**, performed by the source language author(s). Authors use the Controlled Language Checker (CLC) tool for vocabulary and grammar checking on each document they produce. The KANTOO Analyzer is used as a background server which handles individual check requests.
- **Batch document translation**, performed as part of the document production workflow. The KANTOO Analyzer and Generator are utilized as standalone batch servers.
- **Knowledge creation and update**, performed by the domain and language experts. The Knowledge Maintenance Tool (KMT) is used by system developers to edit grammars, structural mapping rules, and other rule-based knowledge in the system.
- **Source terminology creation and update**, performed by domain experts. The Lexical Maintenance Tool (LMT) is used to maintain source terminology in a relational database structure.
- **Target terminology creation and update**, performed by domain translators. The Language Translations Database (LTD) tool is used by translators to create target translations of new source terminology.

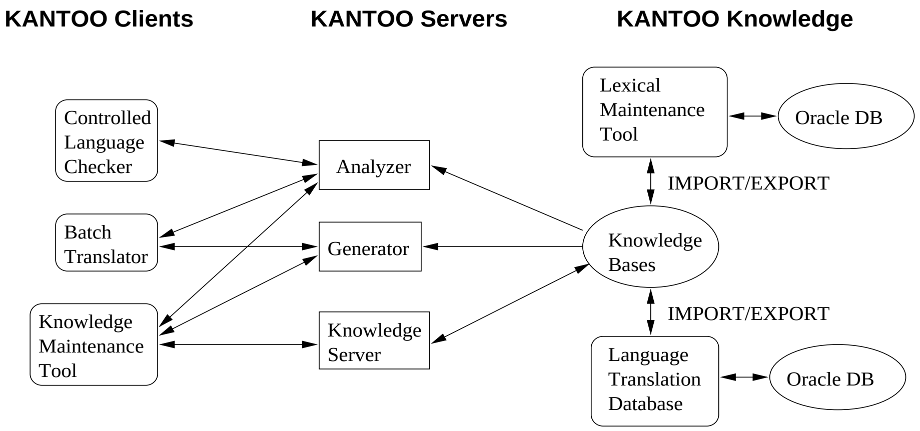

Fig. 1. KANTOO Architecture.

The KANTOO architecture is scalable; several domains, languages, and versions of their knowledge sources can be maintained and executed in parallel. The PC delivery format of the LTD and LMT allow those tools to be used by third-party translations vendors to develop terminology resources. These tools are in daily use at an industrial document production facility [1] for Spanish, French, and German.

2 KANTOO Modules

- **Analyzer**. The Analyzer module performs tokenization, morphological processing, lexical lookup, syntactic parsing with a unification grammar, and semantic interpretation, yielding one or more interlingua expressions for each valid input sentence (or a diagnostic message for invalid sentences). The same Analyzer server can be used simultaneously by the CLC, Batch Translator and KMT[1].
- **Generator**. The Generator module performs lexical selection, structural mapping, syntactic generation, and morphological realization for a particular target language. The same Generator executable can be loaded with different knowledge bases for different languages. The same Generator server can be used by the Batch Translator an KMT in parallel.
- **Lexical Maintenance Tool (LMT)**. The Lexical Maintenance Tool (LMT) is implemented as an Oracle database and Forms application which helps

[1] Space limitations preclude a discussion of a) the Controlled Language Checker, which has been discussed at length in [3], and b) the Batch Translator, which is a simple piece of driver code that uses the KANTOO servers to translate entire documents.

users to create, modify, and navigate through large numbers of lexical entries. The LMT brings together the various kinds of lexical entries used in NLP development, including words, phrases, and specialized entries such as acronyms, abbreviations, and units of measure.
- **Language Translation Database (LTD)**. The LTD is the target language counterpart to the LMT, and is also implemented using Oracle and Forms. The LTD includes productivity enhancements which provide the translator with partial draft translations taken from similar translated terms.
- **Knowledge Maintenance Tool (KMT) and Knowledge Server**. The Knowledge Maintenance Tool (KMT) is a graphical user interface which allows developers to test their knowledge changes in the context of a complete working system. Users can trace or edit individual rules or categories of rules. The KMT operates in conjunction with the Knowledge Server, which provides distributed network access to a version-controlled repository of KANTOO knowledge sources.

3 Knowledge Update in KANTOO

There are two main types of knowledge update in KANTOO: a) terminology updates, which include both source and target language vocabulary; and b) knowledge base updates, which include enhancments and bug fixes made to source and target grammars, mapping rules, etc. to improve translation coverage and quality.

- **Terminology Updates**. When a new version of the source language terminology is released, the contents of the LMT are synchronized with the contents of the LTD. Both databases share a virtual data model, and use the same primary key; the synchronization process ensures that each target language database includes entries for all the new (untranslated) terminology. The individual databases are then distributed to translators, who provide translations for the new terms. Both the LMT and LTD databases are then exported to the machine-readable lexicon format used by KANTOO. Once a set of new lexicons have been created and tested, they are integrated into the production workflow by updating the production knowledge repository. The KANTOO analyzer and generator servers automatically incorporate these knowledge updates when they are restarted.
- **Knowledge Base Updates**. A variety of rule-based knowledge sources must be maintained in the KANTOO system. Chief among them are the syntactic grammars for the source and target languages. The biggest challenges for updating rule-based knowledge sources effectively rest in the potential complexity of the debug/test cycle. Changing a particular rule might result in widespread changes in grammar coverage, or regressive failures. The Knowledge Maintenance Tool (KMT) is used by the developer to test individual updates, with recourse to full regression testing on various reference corpora. All changes to the knowledge are managed under explicit version

control, so that it is straightforward to synchronize the knowledge sources for different releases. The KMT also includes an interactive tracing and debugging environment which utilizes the KANTOO analyzer and generator servers.

4 Current Status and Future Work

KANTOO is implemented in C++ (Analyzer, Generator, Knowledge Server), Java (KMT) and Oracle/Forms (LMT, LTD). KANTOO has been deployed under AIX and Linux, and is currently being tested under Windows NT. The flexibility of the KANTOO client-server architecture supports distributed, parallel development of new applications and robust, scalable deployments. Our current research focuses on the issues related to deploying the KANTOO architecture in an environment where document authoring and document translation are performed by third-party vendors external to the customer site. This architecture is particularly well-suited for the deployment of authoring and translation as distributed internet services, available over the network 24 hours a day.

5 Acknowledgements

We would like to acknowledge David Svoboda and Michael Duggan for their work on the KANTOO Analyzer and Generator; Anna Maria Berta for her work on LTD and LMT; and David Svoboda, Michael Duggan and Paul Placeway for their work on KMT. We also would like to thank Kathy Baker, Margarida Bolzani, Violetta Cavalli-Sforza, Peter Cramer, Eric Crestan, Krzysztof Czuba, Enrique Torrejon, and Dieter Waeltermann for their work on the development of different KANTOO applications.

References

1. 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)
2. 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)
3. Mitamura, T. and Nyberg, E.: Controlled English for Knowledge-Based MT: Experience with the KANT System. In: Proceedings of TMI-95 (1995)
4. Nyberg, E. and Mitamura, T.: The KANT System: Fast, Accurate, High-Quality Translation in Practical Domains. In: Proceedings of COLING-92 (1992)