TermBuilder: A Lexical Knowledge Acquisition Tool for the Logos Machine Translation System

Brigitte Orliac
Logos Corporation
100 Enterprise Drive, Suite 501
Rockaway, NJ 07866

Kutz Arrieta
Logos Corporation
601 South Washington Street, Apt. 301
Seattle, WA 98104

Abstract

Logos 8, the next generation of the Logos Machine Translation (MT) system, is a client server application, which realizes the latest advances in system design and architecture. A multi-user, networkable application, Logos 8 allows Internet or Intranet use of its applications with client interfaces that communicate with dictionaries and translation servers through a common gateway. The new Logos 8 technology is based on a relational database for storage and organization of the lexical data. In this paper, we present TermBuilder, the Lexical Knowledge Acquisition tool developed for Logos 8. The new automatic coding functionality within TermBuilder is significantly improving the process of acquiring new lexicons for MT and other applications.

1 Introduction

Machine translation may be succinctly defined as the mapping of one language into another by electronic means. Mapping between two languages can be "direct" or it can be achieved by means of an intermediate representation, in a transfer-based or interlingua-based approach. In the transfer-based approach, the source language is mapped into an abstract representation that retains language-specific information (the result of analysis). Bilingual modules convert the source language representations into equivalent target language representations. These are in turn input to a generation module.

In the interlingua-based approach, the source language is mapped into one or more language-neutral representations (such as an ontology or a knowledge base) from which the target language is generated.

In its current incarnation, the Logos MT system is a hybrid system which combines features of both rule-based and interlingua-based systems. In the next section, we present the basic characteristics of the new Logos MT system, hereafter referred to as Logos 8. In section 3, we discuss the major issues surrounding the design, maintenance and development of the lexical module in linguistic applications. In section 4, we introduce TermBuilder, the Lexical Knowledge Acquisition tool developed for Logos 8. We conclude the presentation of TermBuilder by listing remaining issues and future development items.

2 Logos 8: the next generation

2.1 Logos 8 System Architecture

Logos 8 represents the next generation of Logos language products. Logos 8 is a client server application, which realizes the latest advances in system design and architecture. A multi-user, networkable application, Logos 8 allows Internet or Intranet use of its applications with client interfaces that communicate with dictionaries and translation servers through a common gateway. The Logos client interfaces and gateway are written in Java to run on all platforms while the translation server is written in C++. The three client interfaces are:

- Translation Client for the submission and retrieval of terminology extraction and translation jobs
- TermBuilder for terminology management
- Administration Client for user registration and permissions

Logos 8 also features a translation memory component, the Logos Translation Memory (LTM), which is based on IBM's Translation Manager and integrated with the Translation Client.

Another development realized in Logos 8 is the conversion of the old lexicon file management system to a Relational Database Management System (RDBMS). The current version uses Oracle as the RDBMS for storing and maintaining lexical data. Logos 8 runs on a Windows NT platform. Plans for porting the application to Unix Solaris have already

been made. Figure 1 represents the basic architecture of Logos 8[1].

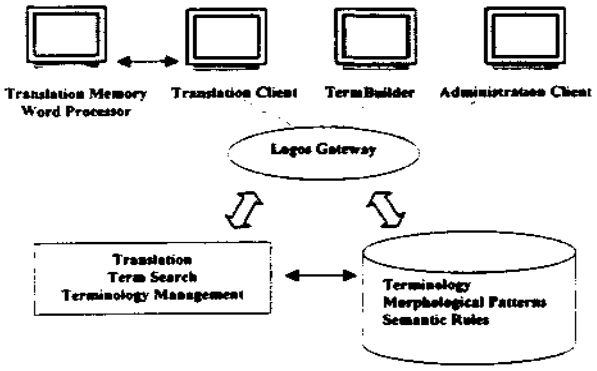

Fig. 1 Logos 8 System Architecture

2.2 Linguistic Model

Natural language in the Logos system is represented as an abstract language with classes or categories that integrate semantic and syntactic properties of words, the Semantico-Syntactic Abstraction Language (SAL). SAL categories exist for all the traditional parts of speech. With over a thousand categories, the SAL ontology is a rich semantico-syntactic hierarchy consisting of four levels of abstraction: a syntactic level (word class) and three concept abstraction levels referred to as superset, set and subset. Each word in the lexicon is classified according to the four levels of the SAL hierarchy. Figure 2 shows a fragment of the SAL ontology for class 1 words (nouns).

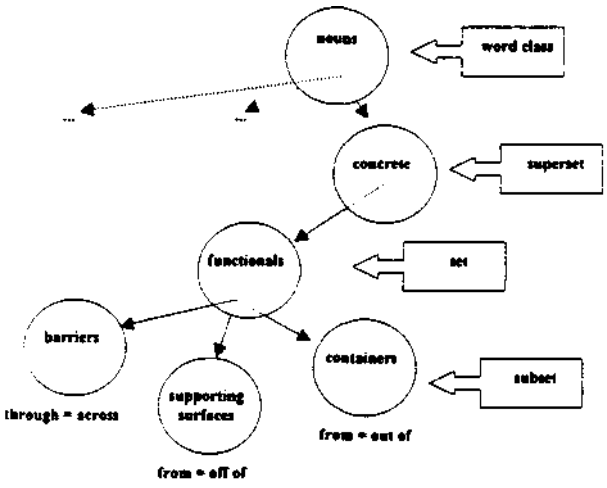

Fig. 2 SAL Ontology for Nouns

[1] The he Logos 8 server will run on Windows NT 4.0, with 512 MB of memory. A 400 MHz Pentium II or higher is recommended. 1.5 GB of disk space should be reserved for the Logos 8 data installation and expansion (this includes the space required for the RDBMS software). The above configuration should translate 40,000 words per hour.

A classification scheme for all words in the lexicon. SAL is an actual language into which natural language is mapped at the outset of the translation process, in dictionary lookup. Once it has been transformed into a series of SAL elements, the source language sentence is matched against a set of linguistic patterns (rules) in the semantic and syntactic rule bases. The linguistic pattern rules in the semantic and syntactic rule bases represent, at various levels of abstraction, semantico-syntactic fragments of the source language environment. When activated by an input vector (the source sentence SAL elements), they interact on the passed elements, incrementally determining the structure and meaning of the source language sentence and constructing an equivalent sentence in the target language.

Translation in the Logos system is performed incrementally, in the six modules of the Translation Server. Specific parsing, transfer or generation tasks assigned to each module are:

- resolve homograph ambiguities and segment the sentence into clauses (RES)
- create the appropriate nodes of a bottom-up parse (Parse 1-4). Parse 1 and 2 are specialized for the lower-level nodes (NPs) while Parse 3 and 4 are specialized for the higher-level nodes (VPs and S).
- expand the nodes in each module into the appropriate phrase structures of the target language (Tran 1-4)
- generate the target language sentence

Figure 3 represents the implementation of the Logos linguistic model in the Logos 8 Translation Server.

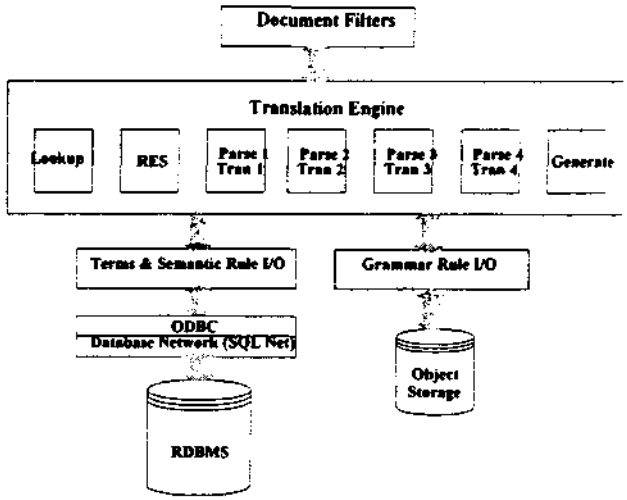

Fig. 3 Logos 8 Translation Server

3 The role of the Lexicon in Linguistic Applications

The lexicon plays a central role in a machine translation system. In some systems, many linguistic operations are generated from the lexicon itself. Dorr (1989, 1993), for example, presents a system in which the lexicon plays an important role in the generation of the interlingua representation. Another example is the encoding of the Lexical Conceptual Structure (LCS) of lexical entries (Jackendoff 1983, 1990).

A fundamental assumption underlies these approaches: that a great deal of the semantic and syntactic information is encoded in the lexicon. While we agree that information such as LCS belongs in the lexicon, we believe that the devices that make use of this information belong in the grammar. In the Logos system, the lexicon is primarily a data repository for lexical entries and their attributes, accessible from the syntactic and semantic rule bases.

Differences in the role assigned to the lexicon in academic and commercial systems reflect a more fundamental divergence: academic systems assume that the translation system is going to "get it right" the first time around while commercial systems tend to put more emphasis on the need for expansion and customization.

Lexical knowledge in the Logos system is organized in a transparent, hierarchical, relationship-based model. Lexical entries are represented as sets of attributes or properties organized in the different tables of the Logos database. The tables are linked in a parent-child relationship, with each table at a different level of linguistic representation. The major dictionary tables are:

- The Word/Phrase table which contains the character string of the word or phrase and other related data (word type, head word, etc.). All strings, irrespective of their function as source or target words, are stored in this table.
- The Morphology table which contains for each entry in Word/Phrase the morphological-syntactic data (word class, gender, inflection pattern, etc.)
- The Meaning table which contains the SAL and subject area attributes
- The Transfer table which contains a pointer to the translation (to a Word/Phrase record and dependent tables in the target language)

Each record (entry) in the Word/Phrase table is linked to one or more records in the Morphology table (homographs such as *break* will have several records, a noun, an intransitive verb and a transitive verb). Each source language record in the Morphology table is linked to one or more records in the Meaning table (one for each identified meaning of the part of speech in question). Finally, each record in the Meaning table is linked to a record in the Transfer table (a pointer to a target language word and its part of speech).

Lexical databases need to be expanded constantly, while preserving all the necessary information and enforcing all prescribed standards (the latter requirement being particularly critical for an MT system). The need to address these requirements largely motivated the conversion of the old lexicon file management system to a relational model. Some of the advantages of an RDBMS are:

- ease of maintenance (a relational model eliminates data redundancy, allowing for single analysis of entry).
- flexibility (model is less static and more expandable. New attributes (fields) can be added as required)
- data integrity via relational constraints
- accessibility (data can be accessed and manipulated more easily)
- scalability (an RDBMS can be scaled to both small and large corporate environments)
- reusability (analyzed target data can be reused to create the lexicon for a new source)
- portability to other platforms

The advantages listed above do not address the problem of acquisition. How does one acquire new data? TermBuilder is the tool that we created to solve this problem.

4 TermBuilder

TermBuilder offers a solution to the problem of the acquisition of new lexical data. But it only offers a partial solution. The problem of lexical acquisition is twofold. On the one hand, there is the problem of finding sources for new data and reusing parts or all of the information found in existing sources. On the other hand, one has to address issues of performance, robustness and accuracy in acquiring the new data. Most of the recent research in the field has focused on the issue of sources for new data and automatic acquisition from existing electronic dictionaries (Farwell, Guthrie and Wilks (1994)).

TermBuilder was designed first to address issues of performance (the old terminology management tool was too slow, the acquisition process required too much knowledge from the user). As the developers of the tool were addressing issues of performance and ease of use, less emphasis was placed on improving the methods for finding and exploiting existing databanks. With linguistic development moving into new areas (the addition of an English-Portuguese language pair), new methods for reusing existing data are being developed, based on the enhanced functionality within TermBuilder.

For the moment, lexical acquisition in TermBuilder takes as its source either documents or existing bilingual or multilingual glossaries. To be imported automatically into the Logos database, existing glossaries need only contain the list of paired terms (source and target), the terms subject area, the source term word class and, where applicable, the source and target noun gender.

4.1 Acquisition of New Data

As mentioned above, lexical entries can be imported into TermBuilder from a document (written in the source language) or from a bilingual glossary.

In the first case, terminology is identified in the source document by executing a Term Search. Term Search searches the document for unfound terms (single words) and candidate terms (noun phrases). Term Search also reports the terms found in the lexicon under specified subject areas. To identify candidate noun phrases. Term Search passes the document through the noun phrase parser of the translation system. Transfers for the phrases are generated based on the target phrase structures produced by the Tran 1 module and are submitted for user review.

The Term Search report is a tab-delimited report where information about each unfound (and found) term is organized in columns that correspond to key attributes of Logos lexical entries: Source Language, Source Term, Source Word Class, Source Gender, Target Language, Target Term, Target Gender, Subject Area. For each reported term, context information is also provided in the form of one example sentence. The Term Search report can be opened in TermBuilder where users can review, complete and finally import their new terminology into the Logos database.

4.2 AutoCode

To facilitate the acquisition of new terminology into the Logos database, we developed an automatic coding or AutoCode functionality. This functionality is one of the most attractive features of TermBuilder. AutoCode supports automatic coding of nouns, verbs, adjectives and adverbs, setting all attributes of the new entry in the Logos database tables. Autocoding in TermBuilder goes through a sequence of processing steps identified below.

The first step is dictionary lookup (similar to dictionary lookup in translation). AutoCode matches the new source and target entries in the database to find already coded Logos entries within a related subject area and reuse them for coding the new entries.

If the dictionary search fails, AutoCode analyzes the unfound noun, verb, adverb or adjective phrase to identify the head word and all inflected modifiers (adjectives and/or noun modifiers).

Compound analysis developed for the dictionary lookup module of the translation system is used within TermBuilder to help identify the head word of a German compound. A simpler string matching logic is used to analyze English compounds.

After the head word of the unfound compound or phrase has been identified, AutoCode assigns the SAL categories. To do so, AutoCode matches the head word of the unfound noun phrase against a list of noun meanings stored for that word in the database. The noun meaning list contains all known meanings of common English and German nouns (and their associated SAL)[2].

Priority ordering within a given noun meaning list is used to automatically determine the meaning of the noun being processed. Upon selection of a meaning. the SAL categories of the selected meaning are retrieved and assigned to the unfound noun phrase When no noun meaning list exists for the head word of an unfound noun or noun phrase, AutoCode assigns default SAL categories. Default SAL categories are also used for unfound adverbs or adjectives (unfound source verbs are not supported currently).

Finally, AutoCode assigns an inflection pattern to the new source and target words, creates the full forms of all inflecting source words and inserts the entries in the Logos database.

4.3 Beyond MT

TermBuilder is currently used in the context of the Logos MT system, but it could also be used in other applications. This extensibility comes mostly from the fact that, with TermBuilder, lexical data can be exported to or imported from different dictionary structures. In addition, given the ease of extraction of the data and of their attributes, lexical entries contained in the Logos database can be used for applications other than MT (authoring tools, etc.).

5 Remaining Issues

There are limitations with TermBuilder. Some of these limitations are inherent to the Logos MT system and designed to ensure the quality of the translated output. For example, TermBuilder will not allow users to update "protected" words (closed class words) or "unknown" or unfound verbs (all verbs are coded manually by Logos linguists). Another limitation concerns the automatic acquisition of acronyms, abbreviations and proper names. These entries have to be done "manually" (in the Add Entry functionality of TermBuilder) as AutoCode cannot assign accurate SAL categories to acronyms, abbreviations and proper nouns (remember that the prompt table contains the known meanings of common nouns only).

[2] To develop AutoCode, a complete revision of the noun meaning data was necessary. During revision, we also expanded the data coverage to include noun meanings which have emerged in recent years.

"Unknown" or unfound entries currently receive default SAL categories. We plan to refine the automatic coding of unfound noun phrases (exploiting available SAL information of the adjective or noun modifier). We are also exploring methods for automatically extracting words and linguistic information from electronic dictionaries (to speed up the development of new language pairs).

6 Conclusion

We have tried to present Logos 8 and TermBuilder in the light of lexical concerns in MT. With the conversion to new system architecture and programming languages and the migration to a relational data model, Logos has clearly shifted its focus from an almost exclusive attention to the syntactic and semantic modules toward the lexicon. Our involvement with issues of representation, coverage and exploitation of lexical data is expected to remain strong in the years to come.

References

Agirre, E. and Rigau, G. (1995). "A Proposal for Word Sense Disambiguation using Conceptual Distance". In cmp-lg.

Bergler, S. (1995). "Generative Lexicon Principles for Machine Translation: A case for Meta-Lexical Structure". In Machine Translation, 9:155-182.

Boguraev, B. and Briscoe, T. (1987). "Large Lexicons for Natural Language Processing: Utilising The Grammar Coding System of LDOCE". In Computational Linguistics, 13-3/4:203-218.

Byrd, Calzolari, Chodorow, Klavans, Neff and Rizk (1987). "Tools and Methods for Computational Lexicology". In Computational Linguistics, 13-3/4:219-240.

Copestake, Briscoe, Vossen, Ageno, Castellon, Ribas, Rigau, Rodriguez and Samiotou (1995). "Acquisition of Lexical Translation Relations from MRDS". In Machine Translation, 9:183-219.

Dorr, B. J. (1989). "Conceptual Basis of the Lexicon in Machine Translation". Parsing Project Working Papers 3. Center for Cognitive Science. MIT Press. Cambridge, Massachusetts.

Dorr, B. J. (1993). "Machine Translation: A view from the Lexicon". MIT Press. Cambridge, Massachusetts.

Dorr, B. J., Garman, J. and Weinberg, A. (1995). "From Syntactic Encoding to Thematic Roles: Building Lexical Entries for Interlingual MT". In Machine Translation, 9:221-295.

Evans, R. and Gazdar, G. (1996). "DATR: A Language for Lexical Knowledge Representation". In Computational Linguistics, 22-2:167-216.

Farwell, D., Guthrie, L. and Wilks, Y. (1994). "Automatically Creating Lexical Entries for ULTRA, a Multilingual MT System". In Machine Translation, 9:128-145.

Fontenelle, T., Adriaens, G. and De Braekeleer, G. (1994). "The Lexical Unit in the Metal® MT System". In Machine Translation, 9:1-19.

Gdaniec, C. and Schmid, P. (1995). "Constituent Shifts in the Logos English-German System". In Proceedings of TMI 95, Leuven.

Hoffmann, E. and Orliac, B. (1999). "Data Exchange between OTELO and Logos". Final report on OTELO Work Package LR 2.3. Logos Corporation.

Hutchins, W. J. and Somers, H. L. (1992). "An Introduction to Machine Translation". Academic Press. London.

Jackendoff, R. S. (1983). "Semantics and Cognition". MIT Press. Cambridge, Massachusetts.

Jackendoff, R. S. (1990). "Semantic Structures". MIT Press. Cambridge, Massachusetts.

Miller, A. (1993). "Nouns in WordNet: A Lexical Inheritance System". Electronic Publication.

Orliac, B. (1998). "The Logos8 System". In Machine Translation and the Information Soup. Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas, AMTA'98, Springer.

Scott, B. E. (1989). "The Logos System". In Proceedings of MT Summit II, Munich.

Scott, B. E. (1977). "Linguistic and Computational Motivations for the Logos Machine Translation System". Technical Report. Logos Corporation.