The Automatic Creation of Lexical Entries for a Multilingual MT System

David Farwell, Louise Guthrie and Yorick Wilks

Computing Research Laboratory
Box 30001
New Mexico State University
Las Cruces, NM 88003-0001

ABSTRACT

In this paper, we describe a method of extracting information from an on-line resource for the construction of lexical entries for a multi-lingual, interlingual MT system (ULTRA). We have been able to automatically generate lexical entries for interlingual concepts corresponding to nouns, verbs, adjectives and adverbs. Although several features of these entries continue to be supplied manually we have greatly decreased the time required to generate each entry and see this as a promising method for the creation of large-scale lexicons.

1. Introduction

For some time, researchers in Computational Linguistics and Natural Language Processing (NLP) have eyed machine-readable dictionaries with interest because they might provide a practical resource for overcoming the "lexical acquisition bottleneck". Many researchers, however, view this problem of lexical acquisition as too difficult to solve at present using a machine-readable dictionary, and the result has been that the focus of much research has shifted to identifying the kind of information needed in NLP lexicons [Atkins, 1990; Miike, 1990; McNaught, 1990; Normier & Nossin, 1990; Nirenburg et al., 1990; Hanks, 1991; Pustejovsky & Bergler, 1990; Warwick, 1990; Kay, 1989], the goal being eventually to create a lexical data base that will allow the creation of a lexicon to be used for processing natural language. While we agree that it is unlikely that the information in machine-readable dictionaries is sufficient for this grand data base of facts that will support NLP as a whole, we are optimistic about making use of the information they do provide to support the creation of lexical entries for specific natural language processing systems. In this paper, we present initial results which are specifically related to extracting information automatically from entries in the *Longman Dictionary of Contemporary English* (LDOCE), in order to construct lexical entries for the ULTRA multilingual machine translation system.

We give an overview ofthe ULTRA Machine Translation System and its lexicon (focusing on the information requirements of its lexical entries), and then discuss the lexical entry construction process. Finally, we offer some suggestions for fully automating the entire process.

2. ULTRA

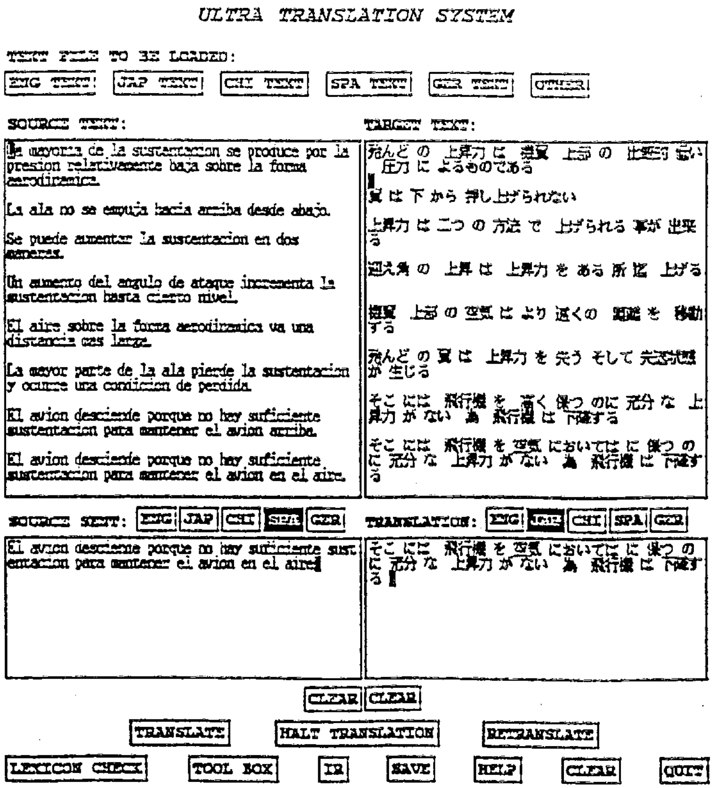
ULTRA (Universal Language TRAnslator) is a multilingual, interlingual machine translation system which currently translates between five languages (Chinese, English, German, Japanese, Spanish) with vocabularies in each language based on about 10,000 word senses. It makes use of recent AI, linguistic and logic programming techniques, and the system's major design criteria are that it be robust and general in purpose, with simple-to-use utilities for customization.

Its special features include:

a multilingual system with a language-independent system of intermediate representations (interlingual representations) for representing expressions as elements of linguistic acts;

- bidirectional Prolog grammars for each language incorporating semantic and pragmatic constraints;
- use of relaxation techniques to provide robustness by giving preferable or ''near miss'' translations;
- access to large machine-readable dictionaries to give rapid scaling up of size and coverage;
- multilingual text editing within X-windows interface for easy interaction and document preparation in specific domains (e.g., business letters, pro-forma memoranda, telexes, parts orders).

Below is a sample screen from the ULTRA system. Each of the Spanish sentences in the "SOURCE TEXT" window have been translated into Japanese. The system has "cut and paste" facilities which allow a sentence from the source text to be moved to the bottom left "SOURCE SENT:" window where it can then be translated by selecting a target language from the choices above the "TRANSLATION" window (bottom right) and choosing the "TRANSLATE" button at the bottom of the screen. The translation then appears in the bottom right "TRANSLATION" window. From there, the translation can then be moved to the "TARGET TEXT" window.



SOURCE TEXT:

La mayoria de la sustentacion se produce por la presion relativamente baja sobre la forma aerodinamica.

La ala no se empuja hacia arriba desde abajo.

Se puede aumentar la sustentacion en dos maneras.

Un aumento del angulo de ataque incrementa la sustentacion hasta cierto nivel.

El aire sobre la forma aerodinamica va una distancia mas larga.

La mayor parte de la ala pierde la sustentacion y ocurre una condicion de perdida.

El avion desciende porque no hay suficiente sustentacion para mantener el avion arriba.

El avion desciende porque no hay suficiente sustentacion para mantener el avion en el aire.

TARGET TEXT:

殆んど の 上昇力 は 機翼 上部 の 比較的 低い 圧力 に よるものである

翼 は 下 から 押し上げられない

上昇力 は 二つ の 方法 で 上げられる 事が 出来る

迎え角 の 上昇 は 上昇力 を ある 所 迄 上げる

機翼 上部 の 空気 に より 遠くの 距離 を 移動 する

殆んど の 翼 は 上昇力 を 失う そして 失速状態 が 生じる

そこ にに 飛行機 を 高く 保つ のに 充分 な 上昇力 が ない 為 飛行機 は 下降する

そこ には 飛行機 を 空気 においてには に 保つ の に 充分 な 上昇力 が ない 為 飛行機 に 下降する

SOURCE SENT:

El avion desciende porque no hay suficiente sustentacion para mantener el avion en el aire

TRANSLATION:

そこ にに 飛行機 を 空気 においてに に 保つ の に 充分 な 上昇力 が ない 為 飛行機 に 下降する

The System of Intermediate Representation

The interlingual representation (IR) has been designed to reflect our assumption that what is universal about language is that it is used to perform acts of communication: asking questions, describing the world, expressing one's thoughts, getting people to do things, warning them not to do things, promising that things will get done and so on. Translation, then, can be viewed as the use of the target language to perform the same act as that which was performed using the source language. The IR serves as the basis for analyzing or for generating expressions as elements of such acts in each of the languages in the translation system.

The representation has been formulated on the basis of an on-going cross-linguistic comparative analysis of hand-generated translations with respect to the kinds of information necessary for selecting the appropriate forms of equivalent expressions in the different languages in the system. We have looked at a number of different types of communication including expository texts, business letters, and e-mail messages and dialogues. This, coupled with the fact that the languages selected for the initial development stage are of different historical and typological background, has led to a solid foundation for developing a flexible and complete descriptive framework.

The Language Components

Each individual language system is independent of all other language systems within ULTRA. Corresponding sentences in different languages must produce the same IR and any specific IR must generate corresponding sentences in the five languages. However, the particular approach to parsing or generation which is used in each of the languages may differ. Each language has its own procedures for associating the expressions of the language with the appropriate IRs. These independent systems communicate by handing each other IRs, and no actual transfer takes place.

Independence of the language-particular systems is of both theoretical and practical interest. Given the required equivalence of the input-output behavior of each of the language systems, this paradigm is excellent for comparing various approaches to parsing or generation for their coverage and efficacy.

A new language may be added to the translation system at any time without unpredictable or negative side effects on the previously developed language systems, or on the system's overall performance.

Furthermore, the addition of any new language system will have the effect of *multiplying* the number of language pairs in the translation system by the number of languages already in the system (having developed an English-Japanese system, we need only develop the Spanish module to have an English-Spanish system and a Japanese-Spanish system, and so forth).

At present, we have developed five prototype language systems for ULTRA. Each system has been implemented in PROLOG as a bidirectional parser/generator. That is to say, in a given language system, the same algorithm is used to do either the analysis or the generation of the expressions of the language.

The system is capable of handling a wide range of phenomena, including compound and complex sentences, relative clauses, complex noun phrases, questions (yes-no and Wh types) and imperatives. There will always be certain classes of non-standard input (e.g. "Where station?") which fall outside the system's normal capabilities and to deal with such irregular input, we are developing a number of techniques which together we call "relaxation". Our assumption is that if a given string or IR cannot be successfully processed even though all the lexical items are available in the system, it should be reprocessed with the various constraints systematically weakened.

ULTRA'S Lexicons

There are two types of entries related to the specification of a lexical item in the ULTRA system: those for intermediate representation (IR) word sense tokens, and those for the words of the individual languages.

Currently, there are eight IR word sense categories including entities (often corresponding to nouns), relations (often corresponding to verbs and adjectives), entity specifiers (often corresponding to determiners), relation specifiers (often corresponding to auxiliaries), case relations (often corresponding to prepositions), proposition specifiers (often corresponding to complementizers), proposition modifiers (often corresponding to sentential adverbials), and conjunctions. Each category is associated with a special set of constraints which ranges in number from one for sentential adverbs, to nine for relations. The number of lexical categories for the individual language lexicons varies from eight to fourteen. There is no simple correspondence between the language-particular lexical categories and the IR categories although the gross relationships stated above appear to hold.

All entries take the general form of simple Prolog unit clauses in (12):

(12) category (Form, F1, F2, ...).

where **F1**, **F2** and so on, are constraints. For language-particular entries, these are generally syntactic constraints associated with an orthographic form, **Form**, such as the gender of a noun, whether a verb is reflexive, and so on. For example, (13) is a simplified and readable version of a Spanish entry for the noun *banco*.

(13) noun (banco, third_singular, masculine, bank4_1).

Similarly, (14) is a Spanish entry for the verb *ingresó*.

(14) verb (ingresó, third_singular, finite, past, simple, indicative, active, deposit1_3).

The final argument represents the IR word sense the Spanish form is used to express. This sense token is associated with a sense definition in LDOCE and is used to index the corresponding IR entry.

For IR entries, the features **F1**, **F2**, and so on, correspond to universal semantic and pragmatic constraints on the word sense, **Form**, such as the classification of an entity as countable or not, the semantic case structure of a relation, and so on. For example the IR entry for **bank4_1** would look something like:

(15) entity (bank4_1, class, countable, institution, abstract_object, economics_banking).

while the IR entry for **deposit1_3** would look like:

(16) relation (deposit1_3, dynamic, placing, agent, patient, human, amount, human, abstract_object, economics_banking).

3. The Automatic Construction of Lexical Items

The work on automating lexical entry has drawn upon extensive research at the Computing Research Laboratory in deriving semantic structures automatically from large machine-readable dictionaries [Slator, 1988; Wilks & Slator, 1989; Guthrie et. al 1990]. Much of the core IR lexicon has been derived from the 72,000 word senses in LDOCE. Codings from the dictionary for such properties as semantic category, semantic preferences and so on have been used, either directly or indirectly, to generate partial specifications of some 10,000 IR tokens for the system.

The partially automated lexical entry process proceeds in three steps: 1) given a sense in LDOCE, an entry is constructed by a process of automatic extraction and formatting of information in the form of a standardized data structure, 2) any remaining unspecified information in that structure is provided interactively, followed by 3) the automatic mapping from the fully specified data structure to the corresponding Prolog facts. Step 3) is very straightforward and will not be described here. Below we give a short description of LDOCE and then discuss the techniques we have used to accomplish steps 1) and 2).

LDOCE

The *Longman Dictionary of Contemporary English* [Procter et al., 1978] is a full-sized dictionary designed for learners of English as a second language. It contains 41,122 headword entries, defined in terms of 72,177 word senses, in machine-readable form (a type-setting tape). With few exceptions, the definitions in LDOCE are stated using a control vocabulary of approximately 2,000 words. The control vocabulary words tend to be highly ambiguous (approximately 17,000 senses are listed in LDOCE for the 2,000 spelling forms).

Both the book and tape versions of LDOCE use a system of grammatical codes of about 110 syntactic (sub)categories which vary in generality. Nouns, for example, may be assigned categories such as *noun*, or *count-noun* or *count-noun-followed-by-infinitive-with-TO*, or *vocative-noun-used-in-direct-address*. The syntactic categories for verbs are particularly extensive and include categories such as *transitive-verb-followed-by-the-infinitive-without-TO*.

In addition, the machine-readable version of LDOCE contains codes which are not found in the book and among them are codes which specify the semantic class of a noun (as one of 34 categories) and the semantic preferences on the complements of verbs and adjectives.

From LDOCE to a Partially Specified Entry

The mapping process from LDOCE to ULTRA word sense entries assumes a particular linguistic context. All the information contained in the LDOCE definition is automatically extracted and used in the appropriate ULTRA specification. For some parts of speech (e.g., nouns), most of the information stored in the interlingual entry can be extracted automatically; for others (e.g., verbs and adjectives), only a portion of the information is available.

For this project we began with a Lisp version of LDOCE, which formats the information from the type-setting tape [Boguraev et al., 1987]. To date, we have extracted information from LDOCE nouns for specifying IR entries for entities, from verbs and adjectives for specifying IR entries for relations, and from adverbs for specifying IR entries for relation modifiers and proposition modifiers. These are the major open class categories of IR word sense tokens and constitute over 95% of the tokens defined thus far. Below we summarize the information required by the categories corresponding to nouns and to verbs (the information which is currently provided automatically is marked by @).

Entities:

@ the sense token indexes a corresponding LDOCE word sense definition,

@ whether it is a class term, the name of an individual, or an anaphoric element,

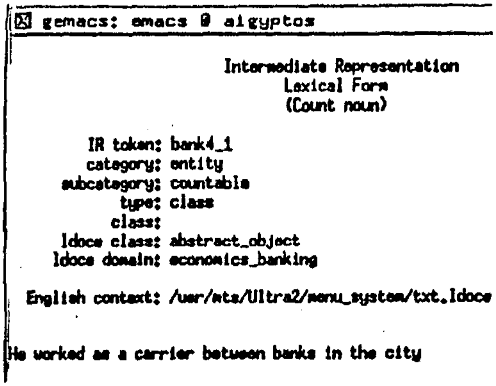
@ whether it is countable or not,

• the semantic class,

@ the LDOCE semantic class,

@ the LDOCE subject domain;

Below is a sample screen of the interactive session for completing the IR lexical entry for one sense of "bank" in LDOCE. The first screen is created automatically and completed manually to produce the second screen.

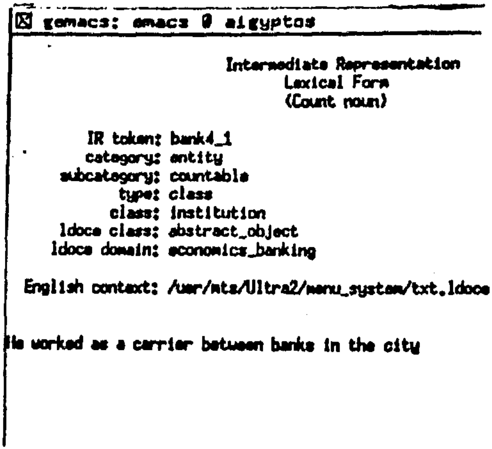


```
                              Intermediate Representation
                                      Lexical Form
                                      (Count noun)

        IR token: bank4_1
        category: entity
     subcategory: countable
            type: class
           class:
     ldoce class: abstract_object
    ldoce domain: economics_banking

 English context: /usr/mts/Ultra2/menu_system/txt.ldoce

He worked as a carrier between banks in the city
```



```
                              Intermediate Representation
                                      Lexical Form
                                      (Count noun)

        IR token: bank4_1
        category: entity
     subcategory: countable
            type: class
           class: institution
     ldoce class: abstract_object
    ldoce domain: economics_banking

 English context: /usr/mts/Ultra2/menu_system/txt.ldoce

He worked as a carrier between banks in the city
```

Note that for entities (nouns) only one feature, described above as "the semantic class," is not provided automatically from LDOCE. This field corresponds to the semantic categories used in ULTRA prior to the use of LDOCE for automatic extraction. These categories were hand crafted, based on surface linguistic phenomena and are used to satisfy the semantic preferences of adjectives and verbs. The automatically created entries for entities contain the LDOCE semantic categories as well, but these will not be used by ULTRA until we have examined the consistency of the LDOCE categories as a basis for semantic preferences.

Relations:

@ the sense token indexes a corresponding LDOCE word sense definition,

- whether it is stative or dynamic,
- the semantic class,

@ the number of case roles,

- the case roles,
- the semantic preferences for the fillers of the case roles,

@ the LDOCE semantic preferences for the fillers of the case roles,

@ the LDOCE subject domain;

In the case of relations, LDOCE does not provide case roles or semantic classes (for verbs), or a direct marking as to whether a verb is stative or dynamic. We have developed a verb hierarchy from LDOCE, based on the genus (hypernym) of a verb definition, and are in the process of disambiguating the terms in this hierarchy. These then will be used as the verb classes for ULTRA's relations. We have been able to extract case role information in some cases [Wilks et al.; 90] from implicit information in Longman's and will include this in the lexical entries. Again the semantic preferences for the fillers of the case roles are those originally used in ULTRA. As in the case of entities above, the LDOCE semantic preferences are also included in the entry for future use.

Extraction is performed by applying a sequence of **flex** programs (a new generation version of the UNIX lexical analyzer utility, **lex**) which transform information from the LDOCE Lisp format into a Lisp association list, the data structure used by the interactive lexical entry interface for the ULTRA system (sample screens appear in the previous secton).

The word senses added to the ULTRA system using these techniques were chosen first on the basis of whether they were exemplified in the dictionary entry, and second, whether they were one of the first three senses of a given homonym (the LDOCE senses are listed in order of frequency of use). Files containing the definitions of all noun, verb, adverb and adjective senses for which there were example sentences were first automatically generated. An additional file containing example sentences tagged by the word sense being exemplified was also created. Next, association lists corresponding to IR entries for each of the word senses were generated. Finally, another procedure was applied which automatically supplied a pointer to the example context in the example sentence file.

4. Approaches to Achieving Full Specification

It was clear at the outset of this project that a great deal of lexical acquisition could be done automatically and we have initiated projects to investigate whether the missing information can be identified automatically through further analysis of the defintions, examples, gramatical categories, etc.

Finally, in order to automate the construction of lexical items fully on the fly during translation, procedures must be defined to select specific senses on the basis of the source language linguistic context of the item being defined. Similarly, procedures must be developed to automatically specify the different language-particular lexical entries (these procedures do exist in English to a limited extent), and these must be adapted to other languages. Finally, techniques for using bilingual dictionaries in the language-specific lexical specification process must be developed.

References

Atkins, B. (1990) The dynamic database, a collaborative methodology for developing a large-scale electronic dictionary. *Proceedings of the International Workshop on Electronic Dictionaries*, Japan Electronic Dictionary Research Institute, Ltd., Oiso, Japan.

Boguraev, B., T. Briscoe, J. Carroll, D. Carter, and C. Grover. (1987) The derivation of a grammatically indexed lexicon from the Longman Dictionary of Contemporary English. *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, pp. 193-200.

Guthrie, L., B. Slator, Y. Wilks and R. Bruce (1990) Is there content in empty heads? *Procedings of the 15th International Conference on Computational Linguistics* (COLING-90) Helsinki, Finland pp. 138-143.

Hanks, P. (1991) The role of dictionaries in language engineering, an Oxford view, Preprint.

Huang, X-M. (1988) Semantic analysis in XTRA, an English--Chinese machine translation system. *Computers and Translation*, 3:1, pp. 101-120.

Jin, W., and R. Simmons. (1986) Symmetric Rules for Translation of English and Chinese. *Computers and Translation*, 1:3, pp. 153-167.

Kay, M. (1989) The concrete lexicon and the abstract dictionary. *Proceedings of the 5th Annual Conference of the UW Centre for the New Oxford English Dictionary*, Oxford, England, pp. 35-41.

Miike, S. (1990) How to define concepts for electronic dictionaries. *Proceedings of the International Workshop on Electronic Dictionaries*, Japan Electronic Dictionary Research Institute, Ltd., Oiso, Japan.

McNaught, J. (1990) Re-usability of lexical and terminological resources: steps towards independence. *Proceedings of the International Workshop on Electronic Dictionaries*, Japan Electronic Dictionary Research Institute, Ltd., Oiso, Japan.

Nagao, M., J-C. Tsujii, and J-C. Nakamura. (1985) The Japanese government project for machine translation. *Computational Linguistics*, 11:2-3, pp. 91-110.

Nirenburg, S., L. Carlson, I. Meyer, and B. Onyshkevych. (1990) Lexicons for KBMT. *Proceedings of the International Workshop on Electronic Dictionaries*, Japan Electronic Dictionary Research Institute, Ltd., Oiso, Japan.

Normier, B. and M. Nossin. (1990) Genelex project: Eureka for linguistic engineering. *Proceedings of the International Workshop on Electronic Dictionaries*, Japan Electronic Dictionary Research Institute, Ltd., Oiso, Japan.

Pereira, F., and D. Warren. (1980) Definite Clause Grammars for language analysis: -a survey of the formalism and a comparison with augmented transition networks. *Artificial Intelligence*, 13, pp. 231-278.

Procter, P., R. Ilson, J. Ayto, et al. (1978) *Longman Dictionary of Contemporary English*. Harlow, UK: Longman Group Limited.

Pustejovsky, J., and S. Bergler. (1987) The acquisition of conceptual structure for the lexicon. *Proceedings of the 6th National Conference on Artificial Intelligence*, pp. 556-570.

Slator, B. (1988) Lexical semantics and Preference Semantics analysis. *Memoranda in Computer and Cognitive Science*, MCCS-88-143, Computing Research Laboratory, New Mexico State University, Las Cruces, New Mexico.

Uszkoreit, H. (1986) Categorial Unification Grammars. Report 66, Center for the Study of Language and Information, Stanford, CA.

Warwick, S. (1990) Automated lexical resources in Europe: a survey. University of Geneva working paper.

Wilks, Y., D. Fass, C. Guo, J. McDonald, T. Plate, B. Slator (1990). Providing Machine Tractable Dictionary Tools. Journal of Machine Translation, 2. Also to appear in Theoretical and Computational Issues in Lexical Semantics , J. Pustejovsky (Ed.)