Structuring a Multilingual Multipurpose Lexical Database Using a Simple Interlingual Approach

Rémi Zajac
rzajac@crl.nmsu.edu

For structuring a Multilingual Multipurpose Lexical Database, we advocate the use of a simple interlingua based on word senses where concepts have no internal structure. This type of interlingua can be used for developing NLP lexicons from Machine-Readable Dictionaries and can serve as the foundation of more elaborated interlingual lexicons.

Background

CRL had and has several multilingual projects concerning multilingual machine translation, multilingual tools for translators and multilingual information retrieval and extraction. The languages concerned include: Arabic, Chinese, English, German, Japanese, Spanish, Russian, and Serbo-Croat. From the breadth of lexical work being pursued at CRL, the need for a multipurpose multilingual database should be obvious. Let me explain more precisely what is meant by multipurpose in the context of the lexical work at CRL. The Mikrokosmos project is a multilingual machine translation project using an interlingua (the "Text-Meaning Representation") linked to an ontology; the Corelli project is a multilingual machine translation project using a glossary-based translation approach and lexical

transfer; the Norm project has built a translator's tool-set including on-line electronic bilingual dictionaries; the information retrieval and extraction projects (part of the Tipster and TREC programs) use bilingual dictionaries and thesauri for generating multilingual queries.

A bare-bones interlingua

In order to build a multilingual multipurpose lexical database with limited resources, a rational choice is to use an interlingua structure which limits the number of mappings between the various languages described in the database. As I will argue thereafter, the use of a bare-bones interlingua, like the one advocated in Sérasset [94a, 94b] does not prevent the definition of lexical transfer relations (for transfer-based MT systems for example) and moreover, it is entirely compatible with more sophisticated versions of interlinguas, such as TMRs.

With drastic constraints on the resources available for building such a database, reuse of existing dictionaries developed at CRL, mostly from Machine-Readable versions of paper Dictionaries (MRDs), is the only approach we can use, and a sensible approach is to use a simplified version of the interlingua defined in the Ultra project. In this project, a concept of the interlingua has a one-to-one correspondence with a word sense of the Longman Dictionary of Contemporary English (LDOCE), and has a structure (which includes for example, the arguments for a predicative concept). In order to accommodate the constraints mentioned above, I advocate two changes to the definition of the interlingua.

It must accommodate various interlingual theories as well as transfer-based relation: the concept of the Corelli interlingua will not have any structure in itself but various theories can be defined and grafted on the interlingua, enhancing the database.

It must accommodate a wide variety of languages and be open to new languages as well as new lexical material: the concepts of the interlingua will not be restricted to the set of word senses from LDOCE, but will be the union of word senses found in all bilingual dictionaries used to build the database.

There are of course well-known problems associated with the proliferation and the management of concepts in this approach, problems that I will qualify, since they are certainly not of conceptual nature, as engineering problems. It must be noted that all interlingual approaches must solve this problem in some way, and they

can, for example, choose to do so by limiting the number of concepts in the interlingua, with a trade-off: augment the complexity of the internal of a concept to be able to represent all sense distinctions in all languages.

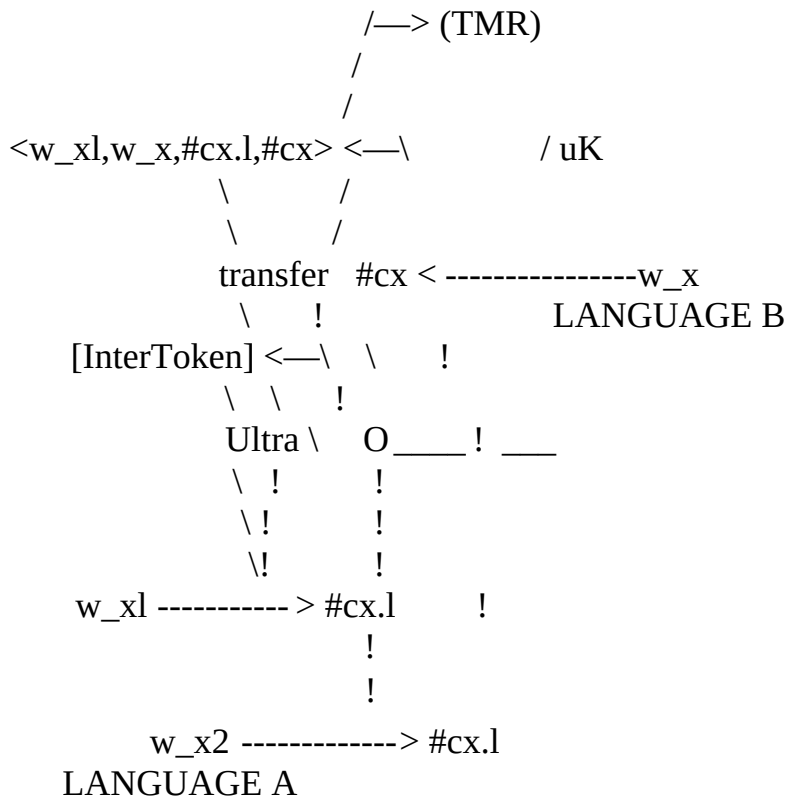

Figure 1 : Translation mismatch and multi-theoretical approach.

Figure 1 shows the relationships between different components of the lexical database for the case of a translation mismatch between a word w_x in language B that can be translated as w_xl or w_x2 in language A. Each word sense has a concept (represented by an arbitrary symbol) in the interlingua: #c_x and two 'sub- concepts' #c_x.l and #c_x.2 (I will come back to the notion of sub-concept or sense refinement in the next section). In the Mikrokosmos approach, each word sense is related to one TMR; in this case, according to the guidelines specified to mapping a sense to a TMR, the predicate of all three TMRs would probably be exactly the same concept of the Ontology (which has a rather different structure than the interlingua structure as presented here), the difference being expressed as a difference in some attribute [Meyer et al., 90, Nirenburg 94, Mahesh 96]. In Ultra, each concept (represented as a Prolog predicate) would correspond to an English word sense and all concepts *have* a translation in each language [Farwell et al., 93]. The 'super-concept', #cx, would simply not exist except in cases of true hyperonymy in English. Within the proposed interlingua structure, assuming we want to use an approach similar to Ultra's, it would be necessary to specify conditions and

transformations on the mapping from one sense (concept) to another since not all concepts are linked to words in all languages. This structure would, however, support more directly a mixed interlingua and transfer-based approach such as the one adopted by EDR [90] which define contrastive relations between two lexical entries by referring to the associated concepts.

Formal properties

From a mathematical point of view, the interlingua has no existence of its own and is no more than a convenient trick to represent in a compact graphical notation, a relation between word senses. In our approach, a word sense has no structure, it is merely a symbol in some set which is defined as the set of word senses in a given language, a convenient way of referring to a lexical sub-entry describing this word sense. Similarly, a concept has no internal structure, it is only a way of relating synonymous word senses between various languages: it simply defines a tuple <t1, t2,-, tn> of word senses t with ti being a word sense in language i. Cases like the one shown in Figure 1 add some interest to this otherwise rather boring structure. To simplify the notations, suppose that we have only 2 languages A and B: Figure 1 pictures the relation defined by the couples <w_xl, w_x> and <w_x2, w_x>, it is a compact graphical representation of the translation relation between these three word senses, factorizing the tuple notation by representing each element of a relation only once and using a disjunctive notation to represent ‘sub-senses' of the interlingua. This view suggests that we can derive simple formal properties on the interlingua from the relational model, for example, that in a given interlingua sub-graph, there must be a link to a word sense in each of the language, otherwise the translation relation is not well-formed.

From a linguistic point of view, however, the interlingua has classically a lattice structure representing hyperonymy, hyponymy and (true) synonymy relationships. If the monolingual parts of the lexical database contain also these relations in the lexical entries, these relationships can be used either for deriving similar relationships in the interlingua or for checking the coherence between the interlingua and the various monolingual dictionaries (all relationships in a given language must also hold -modulo transitivity of relations- in the interlingua). The process of creating the interlingua is then essentially the merging of monolingual lattices of word sense relationships.

Road-map

The construction of a multilingual multipurpose lexical database is not unrelated to the approach of Knight and Luk [94]. The emphasis, however, is not on building an ontology but on defining translation relations between word senses in various languages by pairing these word sense through the mediation of a simple interlingua directly derived from these word senses. This interlingua can then be used for supporting mappings to some ontology.

Using MRDs to build NLP lexicons is now a well understood and well documented process, especially in the initial phases of parsing, restructuring and complementing the dictionaries to build electronic versions [Véronis and Ide 92, Farwell et al., 93, Bauer et al., 94]. This is, however, only a preliminary step even with a bilingual dictionary such as the Collins Spanish-English dictionary. Since we want our bilingual lexicon to be reversible, we need to complement the target side by a monolingual dictionary, e.g., the LDOCE [Sanfilippo et al., 92] or use the reverse version of the dictionary (English-Spanish). Adding new dictionaries should be done with bilingual (or monolingual) dictionaries where one of the languages is already present in the database [Chua et Amat 94, Tanaka and Umemura 94]. This processes clearly involves many steps and a lot of meticulous work that must be carefully planned and for which an appropriate toolkit must be available.

References

Daniel Bauer, Frédérique Segond and Annie Zaenen. 1994. "Enriching an SGML-tagged bilingual dictionary for machine-aided comprehension". Technical Report MLTT-011, Rank Xerox Research Centre, Grenoble, France, October 1994.

Choy-Kim Chua and Salina A. Amat. 1994. "From a bilingual non-electronic dictionary to a ready-to-print bilingual/trilingual electronic dictionary". Proc. of the International Conference on Linguistic Applications, 26-28 July 1994, UTMK-USM, Penang, Malaysia. pp178-191.

EDR. 1990. Proceedings of the International Workshop on Electronic Dictionaries, November 8-9 1990, Oiso, Japan. EDR Technical Report TR-031.

David Farwell, Louise Guthrie and Yorick Wilks. 1993. "Automatically creating lexical entries for ULTRA, a multilingual MT system". Machine Translation, 8(3), pp127-145.

Kevin Knight and Steve K. Luk. 1994. "Building a large-scale knowledge

base for machine translation". Proc. of the 12th National Conference on Artificial Intelligence, AAAI'94.

Kavi Mahesh. 1996. "Ontology Development for Machine Translation: Ideology and Methodology". Memorandum in Computer and Cognitive Science, MCCS-96-292, Computing Research Laboratory, New- Mexico State University, Las Cruces, NM.

I. Meyer, B. Onyshkevych and L. Carlson. 1990. "Lexicographic principles and design for knowledge-based machine translation". Technical report CMT-CMU-90-118, Carnegie Mellon University, August 13, 1990.

Sergei Nirenburg. 1994. "Lexicon Acquisition for NLP: A Consumer Report". In B.T.S Atkins and A. Zampolli (eds.), Computational Approaches to the Lexicon. Clarendon Press, Oxford, UK. pp313-347.

Antonio Sanfilippo, and Victor Poznanski. 1992. "The acquisition of lexical knowledge form combined machine-readable dictionaries". Proc. of the 3rd Conference on Applied Natural Language Processing, 31 March - 3 April 1992, Trento, Italy. pp80-87.

Gilles Sérasset. 1994. "Interlingual lexical organization for multilingual lexical databases in Nadia". Proc. of the 15th International Conference on Computational Linguistics - COLING'94, August 5-9 1994, Kyoto, Japan. pp278-282.

Gilles Sérasset. 1994. "Sublim: un système universel de base lexicales multilingues et Nadia: sa specialisation aux bases lexicales interlingues par acceptions". Ph.D. Dissertation, December 1994, Universite Joseph Fourier, Grenoble, France.

Komati Tankage, and Kyoji Umemura. 1994. "Construction of a bilingual dictionary intermediated by a third language". Proc. of the 15th International Conference on Computational Linguistics - COLING'94, August 5-9 1994, Kyoto, Japan. pp297-303.

Jean Véronis and Nancy Ide. 1992. "A feature-based model for lexical databases". Proc. of the 14th International Conference on Computational Linguistics - COLING'92, August 23-28 1992, Nantes, France. pp588-594.

Rémi Zajac. 1990. "A Relational Approach to Translation". Proceedings of the 3rd International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages, June 1990, Austin, TX, USA.