

## MEANING REPRESENTATION FOR KNOWLEDGE SHARING IN PRACTICAL MACHINE TRANSLATION

Kavi Mahesh and Sergei Nirenburg  
Computing Research Laboratory, New Mexico State University  
mahesh@crl.nmsu.edu, sergei@crl.nmsu.edu

### Abstract

Knowledge-based machine translation can be viewed as the problem of extracting and representing the meaning of a text and generating a translation in a target language using the meaning representation. Meaning extraction requires the integration of information present explicitly in a text with common sense and domain knowledge given to the system. Thus, integrating linguistic knowledge of each language with general world knowledge is a central problem in machine translation, especially when more than two languages are involved. In this article we consider the design of a meaning representation that enables language-specific lexicons to share knowledge with a language-independent world model. We illustrate how the underlying core meaning representation can be enhanced in three different ways to arrive at lexical, ontological, and text meaning representations. The meaning representations presented here have been implemented in the Mikrokosmos machine translation system and used to represent Spanish and Japanese lexicons in addition to a broad-coverage ontological world model.

### 1 Sharing Lexical and World Knowledge

Most human knowledge is represented and communicated via natural languages in spite of relatively recent efforts such as Cyc (Lenat and Guha, 1990) to “computerize” common sense knowledge. Extracting the meaning of a natural language text involves integrating the information present overtly in the text (which is typically incomplete) with the linguistic and world knowledge sources given to the system. The design of such a knowledge based natural language processing (NLP) system is inseparable from issues of sharing knowledge between linguistic and world knowledge sources, and, in turn, sharing knowledge representations between lexical, textual, and ontological meaning representations.

Machine translation (MT), especially when it involves more than two languages, is an NLP problem that provides a particularly interesting context to study knowledge shar-

ing between multiple knowledge bases containing different types of knowledge. In multilingual knowledge based machine translation (KBMT), the meaning of a source text is extracted using both linguistic knowledge of the source language and general world knowledge. The resulting meaning representation is then used to generate translations in one or more target languages. Language generation also requires both world knowledge and knowledge of the target language. The internal representation of the meaning of a text must be independent of particular languages to enable translation to or from any of a set of languages.

A fundamental issue in the design of multilingual MT (and other NLP) systems is how to represent (i) linguistic knowledge for each language, (ii) general world knowledge, and (iii) text meanings. It is obvious that there must be one lexicon for each language since languages differ at least in the words they use. For simplicity we assume that all linguistic knowledge for a language is represented in its lexicon without considering the possibility of sharing linguistic knowledge across languages. The question that remains is how to share general world language with individual lexicons for different languages.

In a unilingual NLP system, linguistic and world knowledge can be combined in a single, monolithic representation. World knowledge can be distributed throughout the single knowledge base and inextricably intertwined with linguistic knowledge. Such a scheme has in fact been employed in a variety of NLP systems (e.g., Birnbaum and Selfridge, 1981; Jurafsky, 1992). If the monolithic design is applied to a multilingual system by representing one knowledge base for each language, there must be both a common meaning representation and sharable world knowledge that are left unspecified but are duplicated for each language.

In our approach, linguistic knowledge is represented in separate lexicons for each language while world knowledge is in a language-independent ontology that is shared by each of the lexicons. Sharing of the ontological world model is enabled by having a core language for meaning representation (MR) that is shared by the lexicons and the ontology and also used to represent text meanings internally. Figure 1 shows the relationships between the various knowledge sources in a multilingual MT situation.

This scheme has been developed over the years in a series of MT projects (Nirenburg, et al, 1992; Nirenburg and Levin, 1992; Onyshkevych and Nirenburg, 1994) and is the basis for our current KBMT project called Mikrokosmos. In Mikrokosmos we have developed a Spanish lexicon of over

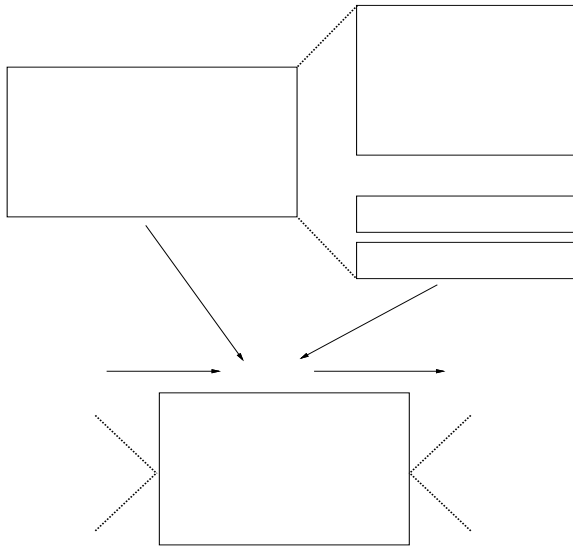


Figure 1: Several Lexicons Sharing a Single Ontology in a Multilingual MT System.

6000 words and a broad-coverage ontology of over 4500 concepts. Acquisition of a Japanese lexicon has also begun recently. Mikrokosmos can currently produce text meaning representations from article-length real-world Spanish texts about company mergers and acquisitions. Although multilingual MT is not new, Mikrokosmos can be seen to be a truly large-scale project in semantic analysis when the richness and coverage of its meaning representations (with respect to a single word or text) are considered together with the nontrivial sizes of its knowledge bases.

The methodology for designing knowledge representations followed in our work has implications for the study of knowledge sharing in general. We first consider the content of knowledge needed for the entire system and identify a core meaning representation that is to be shared by each of the knowledge bases in the system. This core representation is then enhanced minimally in turn for each type of knowledge (or each individual representation) based on a functional analysis of the requirements of the overall task. In this paper, we present the design of the core meaning representation (MR) and illustrate how it is enhanced to arrive at lexical, ontological, and text MRs (Figure 1) to enable knowledge sharing. This work can also be viewed as the beginnings of knowledge sharing between natural language texts in different languages.

## 2 Meaning Representation

Designing a meaning representation for NLP involves determining its content and its representation. We address the issue of content in terms of the nature of primitives used to share knowledge between linguistic and world knowledge representations. The structure of the representation is shown to be based on the needs for composing the primitives in dif-

ferent ways, the expressiveness of which is determined by the machine translation task and its needs for linguistic and world knowledge.

**Semantic Primitives:** A fundamental question in designing knowledge representations for NLP is which and how many primitives to use. Our approach takes an intermediate position between a highly minimalist and a highly excessive number of primitives. The question of primitives is often not addressed in knowledge representation research in other areas of AI and sometimes left unspecified even in NLP systems. Semantic lexicons are sometimes built by introducing a number of primitives as needed for representing word meanings. Neither the set of primitives, nor the taxonomic or other relationships between the primitives is specified in such systems.

**Eleven +/- Two Primitives:** The best example of an extremely minimalist position can be seen in Schank’s (1973) well-known Conceptual Dependency theory (CDs), an ontology of about 11 events.<sup>1</sup> Such a small number of primitives is not practical for building large scale systems that attempt to capture the full richness of meaning that is necessary for MT or other NLP tasks in a domain. When we attempt to decompose complex events such as “a takeover bid for a company” in CDs, the resulting meaning representations will be lengthy, convoluted, and hard to acquire on a large scale. Moreover, they are unsuitable for MT since it is very hard to generate the equivalent word(s) in a target language from such highly decomposed meaning representations.

**Each Word Is a Primitive:** The other extreme position, a popular one in NLP, makes almost every word sense in a natural language a primitive by itself. Examples of such “ontologies” are WordNet (Miller, 1990) and Sensus (Knight and Luk, 1994). In this approach, the large set of primitives is necessarily tied to a particular language (English in the above systems) which may be desirable for MT if the target language was always the same (say English). In a more general multilingual situation, however, the set of primitives must be independent of any natural language. A more compositional meaning representation with a smaller number of language independent primitives is much more practical for constructing large-scale semantic lexicons for multiple languages.

**A Practical Intermediate Position:** We take an intermediate approach and propose a set of primitives that is much bigger than CDs or LCSs but significantly smaller than the typical size (of the order of 50,000) of a “word sense taxonomy” such as WordNet (Miller, 1990). Our experience in Mikrokosmos and its predecessor projects shows that fewer than 10,000 primitives are sufficient for building practical MT systems in a nontrivial domain, such as company mergers and acquisitions.<sup>2</sup> For successful, multilingual MT, such a system must be provided with a rich compositional structure in its meaning representations. The 6000-8000 primitives must also be organized in a highly interconnected ontological network. Using such a scheme, we have built a Spanish lexicon with

<sup>1</sup>Other well-known minimalist approaches include Jackendoff’s (1990) lexical-conceptual structures (LCS). See also (Dorr 1993) See Wilks (1992) or Onyshkevych and Nirenburg (1994) for criticisms of the minimalist approach to meaning representation.

<sup>2</sup>It must be noted that we do not propose to encode only those meanings of words that are in the chosen domain. We in fact encode all meanings of words in a corpus using much less than 10,000 primitives.

over 6000 words that use less than 2500 primitives in their meaning representations.<sup>3</sup>

**Composing Primitives and Expressiveness:** How do we compose the chosen primitives to build lexical representations for word meanings, representations of extra-linguistic world knowledge, and meanings of entire texts? This question is far too often predetermined by one’s methodological commitments. For example, first-order predicate calculus or a frame-based knowledge representation system is commonly chosen without analyzing the functional requirements and practical constraints from the NLP task on the expressiveness of meaning representations. We show how limited expressiveness is an enviable virtue from the practical point of view of building large scale semantic NLP systems. Starting with a minimally compositional representation that is highly limited in its expressiveness, we show what enhancements are necessary in order to match the needs of linguistic, textual, and world knowledge representations.

## 2.1 Core Meaning Representation

The core of our meaning representation (MR) is the set of primitives that are defined as concepts in a language-independent ontology. Since we have fewer primitives than words (or word senses) in a natural language, we must be able to compose two or more primitives to build MRs. In the core MR, primitives can be composed with one another as outlined below:

1. The primitives are first partitioned into free-standing *concepts* (namely, *events* and *objects*), *properties*, and a small set of *literals*. Literals (such as values of color properties: red, blue, etc.) are atomic symbols that cannot be modified further.<sup>4</sup> Properties are not first-class concepts either in that they cannot exist by themselves in the MR. They can only be used to modify other concepts.
2. Properties are further partitioned into *attributes* and *relations*. A property modifying a concept together with the value of that property (i.e., a named value) is called a *slot*. There can be only one value per slot.
3. A multitude of slots can be composed with a concept. Attributes are the names of slots where the values are numbers or literals.
4. Relations are used to represent modifications where the values are other concepts. All conceptual relations are binary and their values must be names of other concepts. A relation always links the concept it is modifying to the single concept that is the value.

The core MR resulting from the above rules for composing primitives is shown in the form of a BNF grammar in Figure 2. It is very limited in expressiveness but constitutes the common foundation for constructing linguistic and world knowledge representations that share a significant amount of language-independent knowledge. In order to show how

---

<sup>3</sup>Construction of a Japanese lexicon using the same set of primitives has also begun recently.

<sup>4</sup>Literals are used to bottom out on unending decompositions of meanings in order to keep MRs simple. Indiscriminate use of literals takes us closer to the word sense approach and is strongly discouraged by our methodological guidelines for representing lexical and world knowledge (Mahesh, 1996; Mahesh and Nirenburg, 1995).

limited this core MR is, we list some commonly used representational apparatus that is missing in the MR:

1. There is no negation operator in the MR.
2. There is no taxonomic organization of primitives or any form of inheritance.
3. It is not possible to compose properties. Properties cannot be attached to other properties.
4. There are no quantification operators. For example, it is not possible to quantify over the set of properties attached to a primitive or represent the existence of an unknown value in a slot.
5. There is no reification operation. It is not possible to raise a slot to the level of a free-standing concept in the MR.
6. It is not possible to make arbitrary (first or higher order) assertions over the representation.
7. The only form of equality test is through simple identity of primitives.

Figure 2: A BNF for the Core Meaning Representation for Knowledge Sharing.

We now take this drastically limited MR and enhance it in necessary ways to formulate suitable representations for lexical, world, and text meanings. This approach guarantees that there will be virtually no need for duplicating any piece of knowledge between the different knowledge bases. Any part of the system’s knowledge that can be expressed in the core MR need be represented only once and can be shared by all the lexical and world knowledge bases. Only those parts that require the following enhancements to the core MR may not be sharable between all the different representations.

## 2.2 Ontological Meaning Representation

In a multilingual MT system, it is best to keep the common world knowledge in a single knowledge base that is shared by the lexicon for each language. We call this common knowledge base the *ontology* and use it to define taxonomic and other relationships between the primitive concepts in the MR. Conceptual relations represent the constraints on potential fillers of slots, rather than actual values (and hence are known in NLP as *selectional restrictions*). In order to represent such conceptual relationships, we need to enhance the MR as follows:

1. We introduce a new type of filler for slots that denotes a constraint on potential values and refer to it as a *sem facet*. A slot now has a property, a sem facet, and a *value facet*. The value facet carries the actual value of a property as formulated originally in MR.
2. To represent constraints on attributes that take numerical values, we allow the sem facets of attribute slots to be a closed or open numerical range (e.g.,  $(0 \leq x \leq 1)$  and  $(x \geq 0)$  respectively).
3. Often, constraints on slots cannot be specified using a single primitive in the sem facet. As a first step, we allow *multiple fillers* in sem facets and define their semantics to be a disjunction of those fillers. For example, the gender property of a human being is constrained to be male or female. Often, such enumeration is prohibitively uneconomical. For example, we could not possibly list all the different types of animals in the world in order to constrain the fillers of a slot to just animals. Hence, we arrange the primitive concepts in a *taxonomic hierarchy*. The new semantics for a sem facet says that the actual value can be any descendant of the constraint in the hierarchy. Since it is very difficult to arrange all concepts in the world in a simple tree, we allow *multiple parents* and make the hierarchy a plex structure.
4. For economy in representing properties of concepts, we introduce *inheritance* so that slots that are defined for parent nodes in the hierarchy need not be repeated in children concepts unless we want to change the fillers. We often need to block inheritance of a particular property to a particular child. Such nonmonotonic inheritance is indicated by a special filler, *\*nothing\** in the particular slot for the child. For example, all animals can be purchased for a price, but not humans (in today’s post-slavery world).
5. Sometimes, a semantic constraint excludes just one or two concepts in an entire subtree in the taxonomy. For economy in representing such constraints, we introduce a limited form of *negation* in sem facets. For example, to constrain the agents of flying to birds, we can say “bird, not penguin, not ostrich” instead of having to list all the other birds.
6. In NLP (and other tasks such as reasoning), it is often useful to know the “typical” value of a slot in addition to the set of all possible values (i.e., the sem constraint). In order to represent such typicality information, we introduce another facet called the *default facet*.
7. Ambiguity resolution and other semantic processes such as the interpretation of non-literal expressions (metaphor, metonymy, etc.) require making selections based on the “semantic closeness” of concepts. In order to support such search operations in the ontology, for each slot that is a relation in any concept, we need to be able to find the inverse relation from the filler to the modified concept. Hence we introduce an *inverse* relation between pairs of relations and arrange the relations also in a hierarchy.

The resulting representation, called Ontological Meaning Representation (OMR) has been used to build an ontology of 4500 concepts covering a wide variety of meanings in the world while paying particular attention to our domain of company mergers and acquisitions.<sup>5</sup> Knowledge represented in

<sup>5</sup>The other type of world knowledge useful for NLP, namely, episodic knowledge of remembered instances of events and objects (such as people, places, organizations, etc.) is represented in a separate knowledge base called the *onomasticon* using the same

the ontology is shared (not duplicated) by lexicons for different languages. It is also shared by the meaning representation for a text.

## 2.3 Lexical Meaning Representation

Meanings of words and other linguistic knowledge necessary for semantic analysis are represented in separate lexicons for each language. Lexical representations must map meanings of words to the syntactic structures in which the word can occur in that language. In addition, lexical representations must be able to modify selectional constraints represented in the ontology to capture the many idiosyncrasies of what is acceptable and what is not in a particular natural language. Lexical meaning representation (LMR) is obtained through the following enhancements to MR:

1. We declare that all properties (i.e., slots) defined for a concept in the ontology are implicitly present in the lexicon whenever a word meaning is represented using the concept name. Thus, ontological knowledge is automatically shared by lexicons. Properties of concepts are repeated in the lexicon only when the values or constraints specified in the ontology must be changed to represent a word meaning.
2. Word meanings are mapped to syntactic structures by binding variables between the syntactic and semantic fields of lexical entries. Such variable binding enables the NLP system to use syntactic guidance in extracting the meaning of a text.
3. In order to further constrain a conceptual relation or an attribute slot specified in the ontology, we borrow the *sem facet* from OMR. To allow the lexicon to relax an ontological constraint, we introduce a fourth facet called *relaxable-to*. These facets allow the lexicon to represent language-specific idiosyncrasies.
4. Since we have fewer primitives than words, meanings of many words are represented by composing more than one primitive concept. Word meanings often map to particular relationships between concepts. Since these concepts can be modified further within the lexical entry, there is a need for referring to the modified concepts in the entry. This is accomplished by introducing variable bindings between the different concepts used in the entry for a word.
5. Word meanings often include various linguistic embellishments such as attitudes, modalities, time, aspect, and so on. These element of meaning cannot be adequately represented in the ontology since they can be attached to almost any concept and since their usage tends to be highly language specific. We view them as additional properties that are not part of the property hierarchy in the ontology.
6. Words often map to sets of objects or events or set-theoretic relations between them. We introduce a set and subset notation to capture such meanings. For example, the word “majority” means a subset of a universal set that is more than half the size of the universal set. Without the expressiveness of the set notation, we would need many more primitives in the system.
7. We also introduce a limited form of negation using a property called *polarity* (with possible values *positive* and *negative*). This is used, for instance, to say that an event did not occur. A limited form of existential quantification is added

OMR representation.

to LMR to represent an existing, but unknown filler of a property using the special symbol \*unknown\*.

Using this representation, we have built a Spanish lexicon with over 6000 entries where meanings of words are represented in LMR by sharing ontological knowledge.

## 2.4 Text Meaning Representation

A text meaning representation (TMR) is made up of particular instances of meanings represented in the ontology and the lexicon. To distinguish between concepts and instances, we introduce an *instantiation operator* that works as follows. An instance has all the properties that its concept has either in the ontology or in the lexicon for which a value is available. Instances only have value facets in their properties. Values are found as follows: any value represented in the lexicon supersedes any value in the ontology. If no value is specified in either the lexicon or the ontology, a default filler from the ontology is the value for the instance.

In addition, the following enhancements are made to MR to get the TMR language:

1. Properties in TMRs often need to refer to chunks bigger than individual instances in the TMR. For example, in "John said that Bill's obsession with guns sent him to prison," what was said was more than a single object or event. In order to support such scoping needs, we introduce a "super structure" called *proposition*. A proposition has a head that must be a concept. In addition, it has a limited set of properties including time, aspect, and attitude. Propositions group word meanings into bigger structural units to represent meanings of entire sentences. A proposition is to TMR what a sentence is to a text.
2. An instance often needs to refer to a particular slot of another instance. We introduce a *reification* operator that raises a slot in an instance to the TMR level by making it an individual instance in the TMR.
3. In order to capture coreferences in a text, we introduce a (reified) *coreference* relation that identifies two or more instances in the TMR as referring to one and the same entity in the world.
4. Certain other linguistic embellishments are also added to the TMR. For example, *time relations* and *quantitative relations* are used to represent the corresponding information.

## 3 Discussion and Conclusions

We have presented the design of linguistic and world knowledge representations for a multilingual KBMT system based on the principles of parsimony, modularity, and compositionality, and the needs of practical methodology. In particular, we showed how having a core meaning representation that is shared by all the knowledge bases enables sharing of general world knowledge among a set of language (or possibly domain) specific lexicons.

Our experience shows that limited expressiveness of representations is a strong virtue from the practical point of view of acquiring large scale knowledge bases. Our restricted representations allow nontechnical acquirers and users to visualize entire knowledge bases statically by browsing through them. Highly expressive representations inevitably turn into massive black boxes that one can ask queries and get answers

from, but cannot browse through to see certain parts independently of the rest of the knowledge base. We believe that our experiences in MT and the above methodological points have significant implications for other application areas as well.

## References

- Birnbaum, L. and Selfridge, M. (1981). Conceptual analysis of natural language. In Schank, R. and Riesbeck, C., editors, *Inside Computer Understanding*, pages 318-353. Lawrence Erlbaum Associates.
- Dorr, B. (1993). *Machine Translation: A View from the Lexicon*. Cambridge, MA: MIT Press.
- Jackendoff, R. (1990). *Semantic Structures*. Cambridge, MA: MIT Press.
- Jurafsky, D. (1992). An on-line computational model of human sentence interpretation. In Proc. 10th National Conf. on AI, AAAI-92, pages 302-308.
- Knight, K. and Luk, S. K. (1994). Building a Large-Scale Knowledge Base for Machine Translation. In Proc. *Twelfth National Conf. on Artificial Intelligence*, (AAAI-94).
- Lenat, D. B. and Guha, R. V. (1990). *Building Large Knowledge-Based Systems*. Reading, MA: Addison-Wesley.
- Mahesh, K. (1996). Ontology development for machine translation: Ideology and methodology. Technical Report MCCS-96-292, Computing Research Laboratory, New Mexico State University, Las Cruces, NM.
- Mahesh, K. and Nirenburg, S. (1995). A situated ontology for practical NLP. In Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing, International Joint Conference on Artificial Intelligence (IJCAI-95), Montreal, Canada, August 1995.
- Miller, G. (1990). WordNet: An on-line lexical database. *International Journal of Lexicography* 3(4) (Special Issue).
- Nirenburg, S. and Levin, L. (1992). Syntax-driven and ontology-driven lexical semantics. In Pustejovsky and Bergler, pp. 5-20.
- Nirenburg, S., Carbonell, J., Tomita, M., and Goodman, K. (1992). *Machine Translation: A Knowledge-Based Approach*. Morgan Kaufmann Publishers, San Mateo, CA.
- Onyshkevych, B. and Nirenburg, S. (1994). The lexicon in the scheme of KBMT things. Memoranda in Computer and Cognitive Science MCCS-94-277. Las Cruces, NM: New Mexico State University. To appear also in *Machine Translation*.
- Schank, R. (1973). Identification of conceptualizations underlying natural language. In *Computer Models of Thought and Language*, Schank, R. and Colby, K., editors, San Francisco: W. H. Freeman Co.
- Wilks, Y. (1992). Review of Ray Jackendoff's *Semantic Structures*, *Computational Linguistics*, vol. 18:1, pp 95-97.