[*First International Conference on Language Resources & Evaluation*, Granada, Spain, 28-30 May 1998]

Evaluation of EuroWordNet- and LCS-Based Lexical Resources for Machine Translation

Bonnie J. Dorr, Antonia Martí, and Irene Castellón

University of Maryland, College Park, MD, USA 20742
and University of Barcelona, Barcelona, Spain 08071
bonnie@umiacs.umd.edu
{amarti,castellon}@lingua.fil.ub.es

Abstract

We evaluate two types of lexical resources with respect to their applicability to interlingual machine translation: (1) a EuroWordNet-based database of bilingual links between Spanish and English words; and (2) a repository of semantically classified verbs with their corresponding Lexical Conceptual Structure (LCS) representations. We examine the utility of these two resources for the task of lexical selection in machine translation. Our approach uses a coarse-grained graph-matching scheme that selects target-language words based on their LCS representation. The final selection of target-language terms is based on a finer-grained linking into the EuroWordNet-based repository.

1 Introduction

Our goal is to use WordNet- and LCS-based information for linking two verbs cross-linguistically during the machine translation process. We select appropriate target-language words based on a graph-matching approach that tests for coverage of the LCS meaning components; we then make a final selection based on access to WordNet links. We describe the development of the two lexical resources that are required for translation: (1) a EuroWordNet-based database of bilingual links between Spanish and English words; and (2) a repository of semantically classified verbs with their corresponding Lexical Conceptual Structure (LCS) representations. We then examine the utility of these two resources for interlingual machine translation.

2 Construction of the EuroWordNet-Based Database

We have developed a lexical repository of bilingual links between Spanish and English verbs. Our two starting points for construction of this database were: (1) Levin's publicly available online index (Levin, 1993) and (2) a morphological Spanish-English lexicon used in a foreign language tutoring system (Dorr, 1997a). We hand-tagged each entry in Levin's book with a synonym set from English WordNet (Miller, 1986; Miller, 1990; Miller and Fellbaum, 1991) and then associated each English verb in Levin with its Spanish counterparts. The transitive closure of these two processes produced a bilingual repository between Spanish and English in which each verb is assigned a semantic class from Levin as well as a concept index into the English-based WordNet. The final database will be incorporated into the Spanish portion of EuroWordNet (Calzolari et al., Forthcoming).

The results of this transitive closure were analyzed previously in (Dorr, Marti, and Castellon, 1997), with the objective of determining the types of mismatches that are likely to arise when we apply this process to additional languages. There were initially 18353 entries (3623 verbs) **in** the bilingual repository. Of these, 3025 entries were hand-verified to be correct and the remaining 15328 entries were modified semi-automatically.[1]

The main modification that was made during the hand verification process was the elimination of incorrect assignments of Spanish verbs to semantic classes due to an association with a high number of polysemous English counterparts. For example, the Spanish verb *escribir* had several English translations: *pen* (as in *John penned a letter to Mary), write* (as in *John wrote Mary a letter),* and *compose* (as in *John composed a letter to Mary).* These English counterparts were mapped automatically into the following semantic classes in the initial bilingual repository:

- pen—9.10 (Pocket Verbs)
- compose—26.4 (Create Verbs)
- compose—26.7 (Performance Verbs)
- write—25.2 (Scribble Verbs)
- write—37.1 (Transfer of Message Verbs)

Of these, only classes 25.2, 26.7, and 37.1 survived hand-revision since 9.10 refers to *pen* in the sense of putting into a pen (not *writing with a pen)* and 26.4 refers to *compose* in the sense of *constructing something* (not *writing something).*

In addition to the elimination of incorrect class assignments, several entries were reclassified into alternative Levin classes. For example, the Spanish verb "acusar" was originally assigned to classes 33 (Judgment Verbs) and 10.6 (Cheat Verbs), but this verb was reassigned to class 13.4.2 (Equip Verbs).

Other entries were deleted due to the rarity of usage and/or disjointness with respect to WordNet concepts, e.g.,

[1] Using automatic and semi-automatic techniques, the entire resource process took 7 months. Adding a new language would take less time since the first step, hand-tagging each Levin-classified verb with a set of WordNet sense, may be reused for other languages. We estimate that it would take at least 2 years to build such a repository from scratch (by human recall and data entry alone), and in such a case, the potential for error would be at least twice as high.

Table 1: Spanish and English WordNet Links For Levin Verbs

English Verb	Levin Class	WordNet Sense Tags	Spanish Verb	Levin Class	WordNet Sense Tags
arrange	class 9.1	Sense 2 (00416049)	clasificar	class 9.1	Sense 2 (00416049)
place/position/put	class 9.1	Sense 1 (00859635)	colocar	class 9.1	Sense 1 (00859635)
remove	class 10.1	Sense 1 (00104355)	borrar	class 10.1	Sense 1 (00104355)
evacuate	class 10.2	Sense 3 (01150129)	desocupar	class 10.2	Sense 3 (01150129)
float	class 11.2	Sense 1 (01069124)	flotar	class 11.2	Sense 1 (01069124)
distribute	class 13.2	Sense 1 (01313552)	repartir	class 13.2	Sense 1 (01313552)
pour	class 26.3	Sense 2 (01184040)	echar	class 26.3	Sense 2 (01184040)

"zapar" (sap). These deletions were (somewhat) balanced off by the addition of new entries—primarily reflexive forms for existing non-reflexive counterparts (e.g., "alarmarse"). The total number of entries in the final Spanish-English bilingual repository is 7319 (3821 verbs). The final format of this repository is illustrated in Table 1.

3 Construction of the LCS-Based Database

We adopt a lexical acquisition approach for Spanish and English based on techniques described in (Dorr, 1997a). We
use Levin's publicly available online index as a starting point.[2] Table 2 shows three broad semantic categories and example verbs along with their associated LCS representations. We have hand-constructed a database containing 191 LCS templates, i.e., one for each verb class in (Levin, 1993). In addition, we have generated LCS templates for 26 additional classes that are *not* included in Levin's system.[3] A full entry in the database includes a semantic class number with a list of possible verbs, a thematic grid, and a LCS template:

(1) **Class 47.8:** adjoin, intersect, meet, touch,...
Thematic Grid: _th_loc
LCS Template:

```
(be loc (thing 2)
     (at loc (thing 2) (thing 11))
     (by !! 26))
```

The semantic class label 47.8 above is taken from Levin's 1993 book (Verbs of Contiguous Location), i.e., the class to which the verb *touch* has been assigned.[4] A verb, together with its semantic class uniquely identifies the word sense, or LCS template, to which the verb refers. The thematic grid (_th_loc) indicates that the verb has two obligatory arguments, a *theme* and a *location.*[5] The ! ! in the LCS Template acts as a wildcard; it will be filled by a lexeme (i.e., a root form of the verb). The resulting form is called a *constant,* i.e., the idiosyncratic part of the meaning that distinguishes among members of a verb class (in the spirit of (Grimshaw, 1993; Levin and Rappaport Hovav, To appear; Pinker, 1989; Talmy, 1985)).[6]

Three inputs are required for acquisition of verb entries: a semantic class, a thematic grid, and a lexeme, which we will henceforth abbreviate as "class/grid/lexeme." The output is a Lisp-like expression corresponding to the LCS representation. An example of input/output for our acquisition procedure is shown here:

(2) **Acquisition of LCS for:** *touch*
Input: 47.8;_th_loc; "touch"
Output:

```
(be loc (* thing 2)
     (at loc (thing 2) (* thing 11
     (by touch 26))
```

Language-specific annotations such as the *-marker in the LCS Output are added to the templates by processing the components of thematic grid specifications, as we will see in more detail next.

An instantiated LCS serves as the interlingua for our machine translation system. For example, the sentence *Mary touched the cat* would have the following LCS representation:

(3) **Mary touched the cat**

```
(be loc (mary)
        (at loc (mary)
        (cat))
        (by touch 26))
```

This representation is the input to our lexical selection procedure which generates the target-language terms, as described in the next section.

4 LCS/WordNet-Based Lexical Selection

One of the main contributions of this work is that it provides the basis for lexical selection in an interlingual machine translation system. Our goal is to use WordNet-based information for linking two verbs cross-linguistically during the machine translation process; this linking will be coupled with an approach to lexical selection based on lexical conceptual structure (LCS). We select appropriate target-language words by producing a candidate set based on a

[2] We focus on building entries for verbs; however, we have approximately 30,000 non-verb entries per language.

[3] Several of these correspond to verbs that take sentential complements (e.g., *coerce).*

[4] Verbs not occurring in Levin's book are also assigned to classes using techniques described in (Dorr and Jones, 1996; Dorr, 1997b).

[5] An underscore (_) designates an obligatory role and a comma (,) designates an optional role.

[6] The ! ! in the Lisp representation corresponds to the angle-bracketed constants in Table 2. For example, the ! ! in a manner position of the Lisp notation corresponds to (MANNER) in Table 2. These constants are linked to the WordNet hierarchy as described in Section 4.

Table 2: Sample Templates Stored in the LCS Database

Category	Verb	Class	Grid	LCS
Location	suspend	9.2	,ag_th,loc()	[CAUSE (X, [BE_{Loc} (Y, [AT_{Loc} (Y, Z)])], [BY ⟨MANNER⟩])]
	touch	47.8	_th_loc	[BE_{Loc} (Y,[AT_{Loc} (Y, Z)], [BY ⟨MANNER⟩])]
Motion	abandon	51.2	_th,src	[GO_{Loc} (Y, [⟨DIRECTION⟩$_{Loc}$ (Y, [AT_{Loc} (Y, Z)])])]
	float	51.3.1	_th,src(),goal()	[GO_{Loc} (Y, [BY ⟨MANNER⟩])]
Placement	adorn	9.8	_ag_th,mod-poss(with)	[CAUSE (X, [GO_{Ident} (Y, [$TOWARD_{Ident}$ (Y, [AT_{Ident} (Y, [⟨STATE⟩$_{Ident}$ ([⟨WITH⟩$_{Poss}$ (*HEAD*, Z)])])])])])]
	spill	9.5	,ag_th	[CAUSE (X, [GO_{Loc} (Y)], [BY ⟨MANNER⟩])]

LCS-based graph-matching approach; we then make a final selection by determining the closeness of LCS constants with concepts in the Spanish-English WordNet database.

Our dual approach provides modularization that parallels the semantic structure / semantic content dichotomy (Levin and Rappaport Hovav, To appear). The graph-matching approach determines the closeness of semantic *structure* between an instantiated LCS (the interlingua) and LCS's stored in the lexical entries of the target language. The final selection determines the closeness of semantic *content* of target-language terms based on access to the WordNet links associated with constants at LCS leaf nodes. We examine the WordNet sense of each constant for which such a link exists; we then use an adaptation of the information-content metric approach by (Resnik, 1995) to select target-language words for those that have no such link.[7]

Consider the following example of translation between English and Spanish:

E: The soldier marched across the field.

S: El soldado marcho atraves el terreno (The soldier marched across the field)
El soldado atraveso el terreno (The soldier crossed the field)
?El soldado atraveso el terreno marchando (The soldier crossed the field marching)

The first of the three target-language sentences is considered to be the most acceptable by native speakers since it contains all relevant information without redundancy. The second sentence is also acceptable, but misses information about marching. The third sentence contains all the relevant information, but is the most awkward. Our MT system will produce all three of these, but will prefer the first two over the second, due to the degree of coverage during graph matching and a preference for minimal concept look-up in WordNet.

[7]The information-content approach computes the distance between "lexemes" associated with the LCS constants. Resnik's system operates on the WordNet taxonomy for nouns only; we are extending his technique for use on the WordNet taxonomy for verbs.

```
E: The soldier marched across the field

LCS for march:
      (go loc
      (thing x)
        (to loc (across loc (thing x) (thing y)))
        (by march/WnSense#l))

IL LCS:
      (go loc
        (thing soldier)
        (to loc
          (across loc
            (thing soldier) (thing field)))
        (by march/WnSense#l))

S: Graph matching:
       LCS for marchar:
          (go loc (thing x)
             (to loc
               (across loc (thing x) (thing y) } )
             (by march))
       LCS for atraveser:
          (go loc (thing x)
            (to loc
              (across loc (thing x) (thing y) ) ) )

S: WordNet Look-up:
       marchar = march/WnSense#l

Final ranked result:
   1. marchar
   2. atraveser
   3. atraveser marchar
```

Figure 1: Lexical Selection Process Using Graph Matching and WordNet Look-up

Figure 1 illustrates the entire process in more detail. The verb *marchar* is selected as the first choice since the graph-matcher retrieves this as an exact match between the IL LCS and the lexical entry for this verb. The second choice is *atraveser* since the graph-matcher finds this as a partial match. Finally, the third choice is *atraveser marchar* since this involves an extra step of performing a WordNet Look-up to fill out the information that is missing from the partial match.

This process is the reverse of the one described in (Dorr, Marti, and Castellon, 1997), where WordNet provided a direct mapping into lexemes which were then selected according to a graph matching scheme. The approach described here is an improvement on this previous approach in that an approximate mapping can be derived for verbs that have no exact correspondence in the target language. The difference can be seen in the following example:

E: He sapped my energy

S: El agoto mi energia[8]

There is no exact translation for *sap* in Spanish, but the graph matching technique pulls out a candidate set of verbs that may be further refined using a distance metric that relates the WordNet sense of *sap* to the Spanish verb *agotar*.

Figure 2 illustrates the entire process in more detail. In this case, the entry for *vaciar* is eliminated after the semantic-closeness metric determines that deplete /WnSense#1 is closer to sap/WnSense#2 than vaciar /WnSense#2.

5 Discussion

We have shown the utility of a EuroWordNet-based database of bilingual links and a repository of semantically classified verb LCS's for the task of lexical selection in interlingual machine translation. We have demonstrated that both resources provide important information that otherwise would not be provided by each resource independently.

Our hypothesis is that synonymous verbs in WordNet are potentially distinguished by their LCS structures; thus we need both inputs—structural (from graph matching) and content (from the WordNet linking) for determining the extent to which the LCS entry for a verb matches the interlingua. That WordNet's hierarchy is shallow for verbs is somewhat balanced by the richness in argument structure provided by the LCS's. The reverse is true for nouns, i.e., WordNet's hierarchy is deep for nouns and their corresponding argument structures are (comparatively) impoverished. Thus, using only one technique or the other (LCS graph matching vs. WordNet linking) cross-categorially would not suffice, in the general case. The combined approach addresses this problem.

We are currently investigating the use of this approach for a large-scale machine-translation effort and also for cross-language information retrieval (See (Dorr and Oard, this volume).)

[8]This phrasing is a bit stilted for Spanish; nevertheless, the main idea is conveyed, and the construction is grammatical.

```
E: He sapped my strength

LCS for sap:
    (cause (thing x)
       (go ident (thing y)
          (toward ident
             (at ident (thing y)
                    (property sap/WnSense#2)))))
IL LCS:
    (cause (thing he)
      (go ident (thing strength)
         (toward ident
           (at ident (thing strength)
              (property sap/WnSense#2))))

S: Graph matching:
     LCS for agotar:
     (cause (thing x)
        (go ident (thingy)
          (toward ident
            (at ident (thing y)
               (property deplete/WnSense#1
    LCS for vaciar:
      (cause (thing x)
         (go ident (thingy)
           (toward ident
              (at ident (thing y)
                 (property   empty/WnSense#2))

S: WordNet Look-up:
       agotar = deplete/WnSense#l
       vaciar = empty/WnSense#2

 Final ranked result:
    1. agotar
```

Figure 2: Lexical Selection Process Using WordNet to eliminate LCS Graph-Matching Possibilities

6 Acknowledgments

The first author has been supported, in part, by Army Research Laboratory contract DAAL01-97-C-0042 and LETTER 11097, NSF PFFIRI-9629108 and Logos Corporation, NSF CNRS INT-9314583, DARPA/ITO Contract N66001-97-C-8540, NSA Contract MDA904-96-C-1250, and Alfred P. Sloan Research Fellowship Award BR3336. The second two authors have been supported, in part, by projects PB-94 0830 of the DGICYT, ITEM TIC-96 1243-C03-02, and EuroWordNet.

References

Calzolari, Nicoletta, Antonia Martí, Horacio Rodriguez, Felisa Verdejo, Piek Vossen, and Yorick Wilks. Forthcoming. Eurowordnet project (title under revision). *Computers and the Humanities.*

Dorr, Bonnie J. 1997a. Large-Scale Acquisition of LCS-Based Lexicons for Foreign Language Tutoring. In *Proceedings of the ACL Fifth Conference on Applied Natural Language Processing (ANLP)*, pages 139-146, Washington, DC.

Dorr, Bonnie J. 1997b. Large-Scale Dictionary Construction for Foreign Language Tutoring and Interlingual Machine Translation. *Machine Translation,* 12(4):271-322.

Dorr, Bonnie J. and Douglas Jones. 1996. Role of Word Sense Disambiguation in Lexical Acquisition: Predicting Semantics from Syntactic Cues. In *Proceedings of the International Conference on Computational Linguistics*, pages 322-333, Copenhagen, Denmark.

Dorr, Bonnie J., Antonia Marti, and Irene Castellon. 1997. Spanish EuroWordNet and LCS-Based Interlingual MT. In *Proceedings of the MT Summit Workshop on Interlinguas in MT*, San Diego, CA, October.

Grimshaw, Jane. 1993. Semantic Structure and Semantic Content in Lexical Representation, unpublished ms., Rutgers University, New Brunswick, NJ.

Levin, Beth. 1993. *English Verb Classes and Alternations: A Preliminary Investigation.* University of Chicago Press, Chicago, IL.

Levin, Beth and Malka Rappaport Hovav. To appear. Building Verb Meanings. In M. Butt and W. Gauder, editors, *The Projection of Arguments: Lexical and Syntactic Constraints.* CSLI.

Miller, George A. 1986. Dictionaries in the Mind. *Language and Cognitive Processes*, 1:171-185.

Miller, George A. 1990. WordNet: An On-Line Lexical Database. *International Journal of Lexicography*, 3:235-312.

Miller, George A. and Christiane Fellbaum. 1991. Semantic Networks of English. In Beth Levin and Steven

Pinker, editors, *Lexical and Conceptual Semantics, Cognition Special Issue.* Elsevier Science Publishers, B.V., Amsterdam, The Netherlands, pages 197-229.

Pinker, Steven. 1989. *Learnability and Cognition: The Acquisition of Argument Structure.* The MIT Press, Cambridge, MA.

Resnik, Philip. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI-95).* (cmp-lg/9511007).

Talmy, Leonard. 1985. Lexicalization Patterns: Semantic Structure in Lexical Forms. In T. Shopen, editor, *Language Typology and Syntactic Description 3: Grammatical Categories and the Lexicon.* University Press, Cambridge, England, pages 57-149.