Lexical Selection for Cross-Language Applications: Combining LCS with WordNet

Bonnie Dorr and Maria Katsova

UMIACS
University of Maryland
College Park, Md 20742
phone: +1 (301) 405-6768
fax: +1 (301) 314-9658
{dorr,katsova}@umiacs.umd.edu
WWW home page: http://umiacs.umd.edu/labs/CLIP

Abstract. This paper describes experiments for testing the power of large-scale resources for lexical selection in machine translation (MT) and cross-language information retrieval (CLIR). We adopt the view that verbs with similar argument structure share certain meaning components, but that those meaning components are more relevant to argument realization than to idiosyncratic verb meaning. We verify this by demonstrating that verbs with similar argument structure as encoded in Lexical Conceptual Structure (LCS) are rarely synonymous in WordNet. We then use the results of this work to guide our implementation of an algorithm for cross-language selection of lexical items, exploiting the strengths of each resource: LCS for semantic structure and WordNet for semantic content. We use the Parka Knowledge-Based System to encode LCS representations and WordNet synonym sets and we implement our lexical-selection algorithm as Parka-based queries into a knowledge base containing both information types.

1 Introduction

This paper describes experiments for testing the power of large-scale resources for lexical selection in machine translation (MT) and cross-language information retrieval (CLIR). We adopt the view that verbs with similar argument structure share certain meaning components [9], but that those meaning components are more relevant to argument realization than to idiosyncratic verb meaning. This distinction mirrors the difference between *semantic structure*, which contributes to structural positioning of arguments, and *semantic content*, which is specific to individual verb meaning.[1]

First, we verify the hypothesis that these two meaning types are distinct by demonstrating that verbs with similar argument structure as encoded in Lexical Conceptual Structure (LCS) [5, 6, 7] are rarely synonymous in WordNet [11, 12, 13]. We then use the results of this work to guide our implementation of

[1] See [10] for more details about the structure/content dichotomy.

an algorithm for cross-language selection of lexical items, exploiting the strengths of each resource: LCS for semantic structure and WordNet for semantic content.

We use the Parka Knowledge-Based System [8, 17] to encode LCS representations and WordNet synonym sets (synsets).[2] Our lexical-selection algorithm is based on Parka-based queries into a knowledge base containing both information types. An input source-language sentence is represented as a LCS; target-language words are then retrieved using LCS-based graph-matching coupled with further refinement by WordNet links.

The advantage of this approach is that it provides a framework for implementing large-scale *event-based* selection using both information types. Event-based selection refers to retrieval on queries that are verb-based clauses (such as 'The soldiers *attacked* the city') or deverbal noun phrases (such as 'The soldier's *attack* on the city'). The benefit to using both LCS and WordNet in event-based retrieval is that the syntactic properties of a word (e.g., that *attack* is a verb in the clause and a noun in the deverbal phrase) are suppressed while more relevant properties are brought into focus: (1) argument structure—that 'soldier' and 'city' are the primary components of the *attack* event; and (2) meaning—that *attack* is closer in meaning to *assault* than to *criticize*. We view the combination of WordNet and LCS as a first step toward evaluating the utility of these two resources for Cross-Language Information Retrieval (CLIR), a large-scale information search task in which the query may be posed in a natural language that is different from that used in the documents [3, 4, 14, 15].

The next section describes our initial experimentation to validate that verbs with similar argument structure are rarely synonymous. Section 3 describes the implementation of a lexical-selection algorithm that exploits this result. Section 4 discusses the impact of the LCS-WordNet combination on the lexical-selection task and describes our future directions.

2 Mono-Lingual and Cross-Lingual Validation of Structure/Content Distinction

We have conducted experiments to verify the hypothesis that verbs with similar argument structure as encoded in Lexical Conceptual Structure (LCS) are rarely synonymous in WordNet. Our experiments were run first mono-lingually and then cross-lingually. An important by-product of these experiments is that, by inducing a reduction in ambiguity for the mono-lingual case, we can achieve more precise results in the cross-lingual case. The idea is that disambiguation of a source-language term reduces the potential "fan-out" in the target language, thus achieving precision close to that of the mono-lingual case (as in traditional single-language IR techniques where no linguistic techniques are used).

We ran experiments with three verbs: *sap*, *walk*, and *close*. We constructed sentences and corresponding input LCSs for each case:

[2] Parka KB provides a very convenient mechanism for studying structural properties of the verbs and to implement fast searching techniques. It also provides a foundation for handling large-scale cross-language resources.

(1) (i) He sapped my strength

```
[CAUSE
  ([Thing HE],
   [GO Ident
      ([Thing STRENGTH],
       [TOWARD Ident
          ([Thing STRENGTH],
           [AT Ident
              ([Thing STRENGTH], [Property SAPPED])])])])]
```

(ii) Florinda walked across the street

```
[GO Loc
   ([Thing FLORINDA],
    [TOWARD Loc
       ([Thing FLORINDA],
        [ACROSS Loc ([Thing FLORINDA], [Thing STREET])])],
    [Manner WALKINGLY])]
```

(iii) He closed the door

```
[CAUSE
  ([Thing HE],
   [GO Ident
      ([Thing DOOR],
       [TOWARD Ident
          ([Thing DOOR],
           [AT Ident
              ([Thing DOOR], [Property CLOSED])])])])]
```

In each of these cases, the semantic structure is encoded in the argument structure itself, e.g., the primitive GO takes as its two arguments a Thing (DOOR) and a Path (TOWARD). The semantic content is encoded as a LCS constant, respectively: SAPPED, WALKINGLY, and CLOSED.

Our experiments were run first on an English database of 54,000 LCS entries that includes verbs, nouns, adjectives, and prepositions. Using a relaxed version of the graph-matching technique described in [3, 4], we ignored constant positions and extracted only those LCSs that structurally matched the input LCS. Consider the verb *sap*. Out of 54,000 LCSs, only 149 match the LCS in (1i). These include verbs like *clean, clear, drain, empty*, etc. We then checked the synonymy relation in WordNet for these graph-matched verbs. The verb *sap*, as used in the LCS above, corresponds to synset 00657546.[3] The only verbs among the 149 graph-matched verbs in this synset are *sap* itself and *drain*. Thus, for this case, we found that semantic-structure/semantic-content overlap occurred in only 2 out of the 149 cases (including *sap* itself).

The full set of results for *sap*, *walk*, and *close* are given in Table 1. Note

[3] The synset numbers are taken from Version 1.5 of WordNet, available at http://www.cogsci.princeton.edu/˜wn. Synset numbers were assigned to LCS templates by hand: each template was human-annotated with as many WordNet synsets as were applicable. (See [1] for more details.)

Verb	Synset(s)	Graph-Matched	Same Synset
sap	00657546	149: clean, clear, drain, empty, erase, reduce,...	2: drain, sap
walk	01086212 01089891 01116106 01086031	272: amble, approach, creep, go, leave, saunter,...	1: walk
close	00772512 00773754	918: collapse, fold, shut, smooth, split,...	2: close, shut

Table 1. Mono-Lingual Generation of Matching Terms

that in each case, the number of graph-matched possibilities is radically reduced (918 down to 2 in the case of *close*), thus supporting our hypothesis that the overlap between semantic structure and semantic content is rare. The two cases where there is more than one overlapping verb (*drain* overlaps with *sap* and *shut* overlaps with *close*) are true cases of syntactic and semantic interchangeability with respect to their usage in the examples given in (1).

In our cross-lingual experiment, we ran the same algorithm on the three LCSs above to produce Spanish target-language equivalents. Our Spanish LCS lexicon has approximately 40,000 LCSs and, as in English, each entry human-annotated with as many WordNet synsets as were applicable.[4] The results in Table 2 show that we were able to restrict the fan-out from Spanish to English words at least as well as in the mono-lingual (English-to-English) case.

We undertook additional experimentation with WordNet to determine if it would be reasonable to produce more target-language candidates, e.g., one link away (hypernymy) from each verb's synset. We found that the candidate set did not grow drastically: one additional term for *sap* (*reducir* = reduce) and one additional term for *close* (*tornar* = change). Further investigation would be required to determine the usefulness of terms generated using other types of links (e.g., hyponymy, troponymy) as well as different distances from the matched target-language candidate. Measures of success would presumably vary on the application: MT would, perhaps, require more refined matching than CLIR. In the next section, we will examine cases where the one-link (hypernym) approach is used to select target-language terms in cases where no synsets match those of the source-language term.

[4] Unlike the English, the Spanish LCS lexicon includes only verbs and nouns (which is the reason for the size discrepancy), but this difference is inconsequential for the event-based experiments reported here. See [2] for more details regarding the annotation of Spanish verbs with WordNet synsets.

Verb	Synset(s)	Graph-Matched	Same Synset
sap	00657546	358: agotar, desaguar, escurrir, evacuar, reducir, vaciar, zapar,...	1: escurrir
walk	01086212 01089891 01116106 01086031	136: andar, caminar, correr, ir, pasear,...	2: andar,caminar
close	00772512 00773754	1554: alterar, cerrar, clausurar, concluir, convertir, disminuir, separar, tapar, virar,...	4: cerrar, clausurar, concluir, tapar

Table 2. Cross-Lingual Generation of Matching Terms

3 Implementation of Lexical Selection Algorithm

Having tested the utility of accessing semantic content independently from semantic structure, we have implemented an algorithm for cross-language selection of lexical items, exploiting the strengths of each resource: LCS for semantic structure and WordNet for semantic content. We use the Parka Knowledge-Based System to encode LCS representations and WordNet synonym sets and we implement our lexical-selection algorithm as Parka-based queries into a knowledge base containing both information types.

Parka is a frame-based knowledge representation that is intended to provide extremely fast inferences and accommodate very large knowledge bases (KBs), on the order of millions of frames. Frames are used to specify categories, instances, and predicates to Parka. Predicates represent relations among entities. The relations being used in our algorithm are binary predicates. We created two tools, one for converting files with LCSs into Parka-based assertions and one for updating the KB (adding new LCSs). We have built Parka KBs for the entire English and Spanish lexicons. We also have transferred all the definitions from the English and Spanish WordNets into Parka-WNet KB.

The basic procedure on the graph-matching (structural) level is the following: Given a composed LCS for a source-language sentence, extract all possible LCSs whose structure covers that of the composed LCS except for the constant position. We implement this procedure by processing the query on each tree level of an LCS representation. Queries are designed to capture only the structural properties of a LCS.

Consider example (1ii) given earlier. The LCS entry for the word *walk* is shown here:[5]

[5] The `[AT]` node is a generic positional primitive that matches any number of other positional primitives such as `ACROSS`, `OVER`, etc.

```
(2) [GO Loc
      ([Thing X],
       [TOWARD Loc ([Thing X], [[AT] Loc ([Thing X], [Thing Y])])],
       [Manner WALKINGLY])]
```

At the highest level, the `GO Loc` node, there are 1059 matching LCSs in the lexicon. For example, the verb *swim* shares this node with *walk*. Moving to the next node level, there are 4399 matches (because all possible matches on two levels are included for each LCS candidate), but the number of possible words has decreased. In general, the algorithm processes the effective query which is optimally constructed for each LCS tree. It extracts all the structural matches of the source LCS on all the tree levels.[6] Finally, the graph-matching procedure extracts the matching target-language words. In the case of *walk*, there are 272 candidates as was indicated in earlier in Table 1.

In order to further reduce this set, we use WordNet as the basis of a more refined lexical selection. For example, suppose we are trying to eliminate *correr* (= *run*) as a target-language candidate. We use WordNet to check for similarity between *runningly* and *walkingly* (or, more precisely, the lexemes themselves: *run* and *walk*). Because *run* is not in any of the synsets containing *walk*, the verb *correr* is ruled out. By contrast, the verbs *andar* and *caminar* are in synsets that include *walk* (both occur in 01086212 and 01086031), so these two verbs are selected as a match.

In addition to cases where target-language terms occur in the appropriate synset(s), we also examined cases where no synsets match those of the source-language term. There are two such cases, one in which there is a LCS that matches exactly (both in structure and in content) and one in which there is no LCS that matches exactly (i.e., the structure matches, but not the content). Thus, including the case where there are matching synsets, there are three cases to consider:

1. If the LCS matches exactly and there are shared synsets, return matching words with shared synsets. For example return *escurrir* for *sap*; *andar, caminar* for *walk*; and *cerrar, clausurar, concluir, tapar* for *close*.
2. If the LCS matches exactly and there are no shared synsets, return words that match exactly. For example, return *fortalecer*, *fortalecerse*, and *confirmar* for *strengthen*.
3. If the LCS does not match in content, return one-link hypernyms of structurally matching words. For example, return *reír* and *reírse* for *giggle*.

In the last case above, we determine the closeness of semantic content using an information-content metric approach (cf., [16]), i.e. selecting those words with the shortest (weighted) distance in WordNet between the mismatched LCS constant

[6] Theoretically, Parka provides utilities to process N-level queries, where N is the depth of the LCS tree. However, due to memory limitations, large-scale application of our algorithm requires that we restrict the number of levels. Thus, at each recursive tree level, we limit our processing to one- or two-level queries.

and the corresponding lexemes. As a first approximation to this, we used the one-link (hypernym) approach to select target-language terms.[7]

Consider the following examples corresponding to the last two cases above:

(3) She strengthened her muscles

```
[CAUSE
  ([Thing SHE],
   [GO Ident
     ([Thing MUSCLE],
      [TOWARD Ident
         ([Thing MUSCLE],
          [AT Ident ([Thing MUSCLE], [Property STRENGTHENED])])])])]
```

(4) Mary giggled at the dog

```
[CAUSE
  ([Thing DOG],
   [GO Perc
     ([Thing MARY],
      [TOWARD Perc
         ([Thing MARY],
          [AT Perc ([Thing MARY], [Thing DOG])])])],
   [Manner GIGGLINGLY])]
```

There are 1554 LCSs in the Spanish lexicon that match the composed LCS structurally in (3). However, none of these correspond to words that share the synsets (00131909 and 00132257) associated with *strengthen* in the composed LCS. Thus, we select only those words whose lexical entry matches the composed LCS exactly, both in structure and in content (i.e., including the constant `STRENGTHENED`). The three words that match exactly are *confirmar* (= confirm), *fortalecer* (= fortify), and *fortalecerse* (= fortify oneself).

In the case of (4), there are 36 LCSs in the Spanish lexicon that match the composed LCS in structure (but not in content). Some examples are: *bufar* (= snort), *cacarear* (= cackle), *gritar* (= howl), *jadear* (= gasp), *reír* (= laugh), *reírse* (= laugh over), and *sonreír* (= smile). However, only *reír* and *reírse* correspond to words that share the synset (00020446) which is a hypernym (one link away) of the set associated with *giggle* in the composed LCS; thus, these two words are selected.

A summary of the last two cases is shown in Table 3

4 Conclusions and Future Work

We have demonstrated that verbs with similar argument structure as encoded in LCS are rarely synonymous in WordNet. We exploit this result in the task

[7] Hypernym links tie a word to its more general counterpart, e.g., *laugh* is a hypernym of *cackle*.

Verb	Synset(s)	Graph-Matched	Same Synset
strengthen	00131909 00132257	1554: alterar, confirmar, fortalecer, fortalecerse, modificar, tornar,... **Exact:** confirmar, fortalecer, fortalecerse	0:—
giggle	00019651	36: bufar, cacarear, gritar, jadear, reír, reírse, sonreír,...	0:— **One link away:** reír reírse

Table 3. Generation of Verbs with no Matching WordNet Synsets

of lexical selection, using LCS graph-matching to determine the closeness of semantic structure (argument structure in the LCS for the events) and WordNet links to determine the closeness of the semantic content (the constant in the LCS for verbs).

The combination of LCS and WordNet allows us to cover a variety of different cases that otherwise would not be handled by either knowledge source alone. In particular, we have shown that there are cases where LCS graph-matching alone is sufficient for selecting target-language terms, e.g., for *strengthen*, where WordNet does not provide a synset-based equivalent in Spanish. We have also shown that there are cases where WordNet is critical to the final selection of target-language terms, e.g., for *walk*, where numerous exactly matched LCSs in Spanish can restricted by a handful of shared WordNet synsets, and for *giggle*, where there are no exactly matched LCSs in Spanish but there exists a small set of related WordNet synsets.

Our future work will generalize the one-link synset matching by integrating a probabilistic technique based on insights from [16], which focuses on nouns. We will implement an analogous information-content metric method for verbs using Parka utilities. We will then extend this combined approach to the task of noun selection. This will involve construction of a probabilistic mapping from Spanish nouns (taken from a Kimmo-based lexicon) and WordNet senses. We expect nouns and verbs to be characteristically opposed in their requirements with respect to the resources we use. In particular, WordNet is hierarchically shallow for *verbs*, but this is counter-balanced by the richness in argument structure provided by the LCSs. In contrast, LCSs are shallow for *nouns*, but this is counter-balanced by the deep hierarchical structure of nouns in WordNet.

Acknowledgments

This work has been supported, in part, by DARPA/ITO Contract N66001-97-C-8540. The first author is also supported by DOD Contract MDA904-96-C-1250, Army Research Laboratory contract DAAL01-97-C-0042, NSF PFF IRI-9629108 and Logos Corporation, NSF CNRS INT-9314583, and Alfred P. Sloan Research Fellowship Award BR3336. We would like to thank members of the CLIP lab for helpful conversations, particularly Jim Hendler, Philip Resnik, Wade Shen, and Scott Thomas.

References

1. Bonnie J. Dorr and Douglas Jones. Acquisition of Semantic Lexicons: Using Word Sense Disambiguation to Improve Precision. In *Proceedings of the Workshop on Breadth and Depth of Semantic Lexicons, 34th Annual Conference of the Association for Computational Linguistics*, pages 42–50, Santa Cruz, CA, 1996.
2. Bonnie J. Dorr, Antonia Marti, and Irene Castellon. Spanish EuroWordNet and LCS-Based Interlingual MT. In *Proceedings of the MT Summit Workshop on Interlinguas in MT*, San Diego, CA, October 1997.
3. Bonnie J. Dorr, Antonia Marti, and Irene Castellon. Evaluation of euro wordnet- and lcs-based lexical resources for machine translation. In *Proceedings of the First International Conference on Language Resources and Evaluation*, Granada, Spain, 1998.
4. Bonnie J. Dorr and Douglas W. Oard. Evaluating resources for query translation in cross-language information retrieval. In *Proceedings of the First International Conference on Language Resources and Evaluation*, Granada, Spain, 1998.
5. Ray Jackendoff. *Semantics and Cognition.* The MIT Press, Cambridge, MA, 1983.
6. Ray Jackendoff. *Semantic Structures.* The MIT Press, Cambridge, MA, 1990.
7. Ray Jackendoff. The Proper Treatment of Measuring Out, Telicity, and Perhaps Even Quantification in English. *Natural Language and Linguistic Theory*, 14:305–354, 1996.
8. Brian Kettler, William Anderson, James Hendler, and Sean Luke. Using the parka parallel knowledge representation system (version 3.2). Technical Report CS TR 3485, UMIACS TR 95-68, ISR TR 95-56, University of Maryland, 1995.
9. Beth Levin. *English Verb Classes and Alternations: A Preliminary Investigation.* University of Chicago Press, Chicago, IL, 1993.
10. Beth Levin and Malka Rappaport Hovav. The Elasticity of Verb Meaning. In *Proceedings of the Tenth Annual Conference of the Israel Association for Theoretical Linguistics and the Workshop on the Syntax-Semantics Interface*, Bar Ilan University, Israel, 1995.
11. George A. Miller. Dictionaries in the Mind. *Language and Cognitive Processes*, 1:171–185, 1986.
12. George A. Miller. WordNet: An On-Line Lexical Database. *International Journal of Lexicography*, 3:235–312, 1990.
13. George A. Miller and Christiane Fellbaum. Semantic Networks of English. In Beth Levin and Steven Pinker, editors, *Lexical and Conceptual Semantics, Cognition Special Issue*, pages 197–229. Elsevier Science Publishers, B.V., Amsterdam, The Netherlands, 1991.
14. Douglas W. Oard. *Multilingual Text Filtering Techniques for High-Volume Broad-Domain Sources.* PhD thesis, University of Maryland, 1996.
15. Douglas W. Oard and Bonnie J. Dorr. Evaluating Cross-Language Text Filtering Effectiveness. In *Proceedings of the Cross-Linguistic Multilingual Information Retrieval Workshop*, pages 8–14, Zurich, Switzerland, 1996.

16. Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of IJCAI-95*, pages 448–453, Montreal, Canada, 1995.
17. Kilian Stoffel, Merwyn Taylor, and James Hendler. Efficient management of very large ontologies. In *Proceedings of AAAI-97*, 1997.

This article was processed using the LaTeX macro package with LLNCS style