Large-Scale Interlingual Machine Translation (NYI) and Development of a Framework for Large-Scale Translation, Tutoring, and Information Filtering (PFF/PECASE)

Bonnie J. Dorr
University of Maryland
Department of Computer Science
and Institute for Advanced Computer Studies
A.V. Williams Building
College Park, MD 20742

1 CONTACT INFORMATION

Email: bonnie@cs.umd.edu
Phone: 301-405-6768
Fax: 301-314-9658

2 OTHER AUTHORS AND AFFILIATIONS:

Nizar Habash, University of Maryland
Douglas Jones, Department of Defense
Maria Katsova, University of Maryland
Douglas Oard, University of Maryland
Mari Olsen, University of Maryland
Scott Thomas, University of Maryland
David Traum, University of Maryland
Clare Voss, Army Research Laboratory

3 RELATED WEB SITE FOR THE PROJECT

http://www.umiacs.umd.edu/~bonnie

4 PROGRAM AREA

2. Speech and Natural Language Understanding.

5 KEYWORDS

- LCS-based machine translation
- multilingual natural language processing

- interlinguas
- lexical acquisition
- cross-language information retrieval
- large scale resources

6 PROJECT SUMMARY

The main goal of the NYI and PFF/PECASE projects (1993–1997,1997–1999) is to investigate the applicability of a lexical-based framework to large-scale natural language processing (NLP) tasks such as interlingual machine translation (MT), foreign language tutoring (FLT), and cross-language information retrieval (CLIR). Specifically, the project focuses on systematizing the relation between syntax and semantics in lexical representations and on constructing lexicons for multilingual NLP applications.

Currently, NLP systems are plagued with problems concerning extensibility to new languages or domains. This problem is addressed by application of a lexical-based framework that accommodates automatic lexicon construction and supports a broader range of cross-linguistic phenomena; the end result is a significant reduction in development time for large-scale NLP systems.

Four results of this research are: (1) a set of novel lexical representations that apply uniformly to languages as diverse as Arabic, English, Chinese, and Spanish; (2) translation and tutoring systems designed to use these novel lexical representations; (3) a general approach to automatic construction of large-scale lexicons for NLP applications; and (4) techniques for cross-language retrieval of texts from a multilingual information collection. The primary benefit of this research is the validation of a common semantic representation that may be used by different researchers to test long-standing hypotheses about computerized translation, tutoring, and information retrieval.

6.1 Lexical Conceptual Structure

An important component of any multilingual NLP system is the representation of meaning for words and sentences in each language. This project has focused on the development and enhancement of a meaning representation called *lexical conceptual structure* (LCS) that supports the development of multilingual NLP systems and automatic acquisition of multilingual lexicons. A set of correspondence rules have been devised to map this representation to its possible syntactic realizations (i.e., the range of possible syntactic structures associated with each word in a sentence). For example, all verbs defined in the semantic class of *Clear Verbs*—clear, clean, drain, empty—are assigned a basic template containing *away_from* as a meaning component; this is associated with a syntactic structure containing a preposition such as *from* or *off* (e.g., *I cleared the dishes off the table*).

While the basic LCS representation accommodates many types of lexical phenomena, more recent work demonstrates that the lexical-semantic representations and correspondence rules alone do not suffice for resolving more complex mismatches [14,15,20,21]. Thus, significant effort has been devoted toward the enhancement of the basic lexical-semantic framework, including the addition of conceptual descriptions for well-defined subclasses of predicates and the identification and formalization of different mechanisms required for resolution of complex linguistic mismatches [12,13,17]. An example is the translation of sentences between languages where there are aspectual distinctions as in the following English-Spanish case:

E: The plane flew across the field.

S: El avión cruzó el campo volando.
'The plane crossed the field flying.'

Despite the different syntactic realizations above, the same underlying semantic representation is assigned to both sentences:

```
(cause (act loc (plane) (flyingly))
       (to loc (plane) (across loc (plane) (field))))
```

The main contribution of this work is the systematization of the relation between syntax (i.e., linguistic structure or 'form') and semantics (i.e., linguistic meaning or 'content') in lexical representations.

6.2 Use of LCS in Implemented Multilingual Systems

The LCS representation is the fundamental meaning representation used in three multilingual systems:

- A foreign language tutoring system (MILT) for Arabic and Spanish [5].
- A large-scale Chinese-English MT system [4,18,19].
- A large-scale cross-language information retrieval system [11,16].

Several computational properties have emerged as a result of incorporating LCS into these systems:

- Verb definitions are systematically partitioned into classes (e.g., *Event* vs. *State*) that enable the specification and application of semantic generalizations; thus, LCS-based lexical entries are efficiently constructed, accessed, and updated [1,2,6,7,8].

- The LCS captures enough meaning to perform accurate lexical selection (i.e., appropriate choice of target-language words) during the generation of an output sentence in MT and FLT systems [3,10].
- LCS principles provide a framework for controlling translation ambiguity in query translation techniques used for cross-language information retrieval applications [9].

6.3 Future Plans

Future plans include the development of routines for extraction of word senses from an interlingual analysis of text in order to add a natural language component to existing information retrieval engines. The main innovation of using an interlingua as the basis for the approach is that it will allow for a variety of languages—both source (i.e., input from a news group) and target (i.e., output into the users' native tongue)—to be incorporated seamlessly into text retrieval systems.

7 PROJECT REFERENCES

[1] Bonnie J. Dorr. 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, 1997.

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

[3] Bonnie J. Dorr, Joseph Garman, and Amy Weinberg. From Syntactic Encodings to Thematic Roles: Building Lexical Entries for Interlingual MT. *Machine Translation*, 9:221–250, 1995.

[4] Bonnie J. Dorr, Nizar Habash, and David Traum. A Thematic Hierarchy for Efficient Generation from Lexical-Conceptal Structure. In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas, AMTA-98, in Lecture Notes in Artificial Intelligence, 1529*, pages 333–343, Langhorne, PA, October 28-31, 1998.

[5] Bonnie J. Dorr, Jim Hendler, Scott Blanksteen, and Barrie Migdaloff. Use of LCS and Discourse for Intelligent Tutoring: On Beyond Syntax. In Melissa Holland, Jonathan Kaplan, and Michelle Sams, editors, *Intelligent Language Tutors: Balancing Theory and Technology*, pages 289–309. Lawrence Erlbaum Associates, Hillsdale, NJ, 1995.

[6] 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.

[7] Bonnie J. Dorr and Douglas Jones. 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, 1996.

[8] Bonnie J. Dorr and Douglas Jones. Use of Syntactic and Semantic Filters for Lexical Acquisition: Using WordNet to Increase Precision. In *Proceedings of the Workshop on Predicative Forms in Natural Language and Lexical Knowledge Bases*, pages 81–88, Toulouse, France, 1996.

[9] Bonnie J. Dorr and Maria Katsova. Lexical Selection for Cross-Language Applications: Combining LCS with WordNet. In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas, AMTA-98, in Lecture Notes in Artificial Intelligence, 1529*, pages 438–447, Langhorne, PA, October 28-31, 1998.

[10] 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*, pages 393–397, Granada, Spain, 1998.

[11] 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*, pages 759–763, Granada, Spain, 1998.

[12] Bonnie J. Dorr and Mari Broman Olsen. Multilingual Generation: The Role of Telicity in Lexical Choice and Syntactic Realization. *Machine Translation*, 11(1–3):37–74, 1996.

[13] Bonnie J. Dorr and Mari Broman Olsen. Deriving Verbal and Compositional Lexical Aspect for NLP Applications. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL-97)*, pages 151–158, Madrid, Spain, July 7-12, 1997.

[14] Bonnie J. Dorr and Martha Palmer. Building a LCS-Based Lexicon from TAGs. In *Proceedings of the AAAI-95 Spring Symposium Series, Representation and Acquisition of Lexical Knowledge: Polysemy, Ambiguity, and Generativity*, pages 33–38, Stanford, CA, March 1995.

[15] Bonnie J. Dorr and Clare Voss. A Multi-Level Approach to Interlingual MT: Defining the Interface between Representational Languages. *International Journal of Expert Systems*, 9(1):15–51, 1996.

[16] Douglas W. Oard and Bonnie J. Dorr. Evaluating Cross-Language Text Retrieval Effectiveness. In Gregory Grefenstette, editor, *Cross-Language Information Retrieval.* Kluwer Academic Publishers, Boston, MA, 1998.

[17] Mari Broman Olsen, Bonnie J. Dorr, and David J. Clark. Using WordNet to Posit Hierarchical Structure in Levin's Verb Classes. In *Proceedings of the Workshop on Interlinguas in MT, MT Summit, New Mexico State University Technical Report MCCS-97-314*, pages 99–110, San Diego, CA, October 1997. Also available as UMIACS-TR-97-85, LAMP-TR-011, CS-TR-3857, University of Maryland.

[18] Mari Broman Olsen, Bonnie J. Dorr, and Scott C. Thomas. Toward Compact Monotonically Compositional Interlingua Using Lexical Aspect. In *Proceedings of the Workshop on Interlinguas in MT, MT Summit, New Mexico State University Technical Report MCCS-97-314*, pages 33–44, San Diego, CA, October 1997. Also available as UMIACS-TR-97-86, LAMP-TR-012, CS-TR-3858, University of Maryland.

[19] Mari Broman Olsen, Bonnie J. Dorr, and Scott C. Thomas. Enhancing Automatic Acquisition of Thematic Structure in a Large-Scale Lexicon for Mandarin Chinese. In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas, AMTA-98, in Lecture Notes in Artificial Intelligence, 1529*, pages 41–50, Langhorne, PA, October 28-31, 1998.

[20] Clare Voss and Bonnie J. Dorr. Toward a Lexicalized Grammar for Interlinguas. *Machine Translation*, 10(1–2):143–184, 1995.

[21] Clare R. Voss, Bonnie J. Dorr, and Mine Ülkü Şencan. The Problem of Lexical Allocation in Interlingua-based Machine Translation of Spatial Expressions. In Patrick Olivier and Klaus-Peter Gapp, editor, *Representation and Processing of Spatial Expressions.* Lawrence Erlbaum, 1998.

8 AREA BACKGROUND

NLP technology provides ways of programming the computer with enough information about language that it can build representations for such tasks as machine translation [23,24,26,32,35], foreign language tutoring [25,30], and multilingual information retrieval [11,33,34]. Currently, NL systems are plagued with problems concerning extensibility, particularly when designers attempt to scale up their systems so that they have broader coverage. The most significant bottleneck in this regard is the construction of machine-tractable lexicons (i.e., large NL databases that relate words to their corresponding meanings) [22,27,28]. To date, designers have been forced to build such lexicons through laborious word-by-word recoding of already existing on-line dictionaries. More recently, researchers have examined linguistically-motivated frameworks—such as that of Levin [31]—which accommodate automatic lexicon construction and cover a wider range of cross-linguistic phenomena. The ultimate goal is to develop techniques that provide a substantial reduction in development time for large-scale NL systems.

9 AREA REFERENCES

[22] Branimir Boguraev and Ted Briscoe, editors. *Computational Lexicography for Natural Language Processing.* Longman, London, 1989.

[23] Bonnie J. Dorr. *Machine Translation: A View from the Lexicon.* The MIT Press, Cambridge, MA, 1993.

[24] Bonnie J. Dorr. Machine Translation Divergences: A Formal Description and Proposed Solution. *Computational Linguistics*, 20(4):597–633, 1994.

[25] Bonnie J. Dorr, Jim Hendler, Scott Blanksteen, and Barrie Migdaloff. Use of LCS and Discourse for Intelligent Tutoring: On Beyond Syntax. In Melissa Holland, Jonathan Kaplan, and Michelle Sams, editors, *Intelligent Language Tutors: Balancing Theory and Technology*, pages 289–309. Lawrence Erlbaum Associates, Hillsdale, NJ, 1995.

[26] Bonnie J. Dorr, Pamela W. Jordan, and John W. Benoit. A Survey of Current Research in Machine Translation. In M. Zelkowitz, editor, *Advances in Computers, Vol. 49.* Academic Press, London, To appear in 1999.

[27] Bonnie J. Dorr and Judith Klavans. Special Issue on Machine Translation: Building Lexicons for Machine Translation I. *Machine Translation*, 9(3–4), 1994/1995.

[28] Bonnie J. Dorr and Judith Klavans. Special Issue on Machine Translation: Building Lexicons for Machine Translation II. *Machine Translation*, 10(1–2), 1995.

[29] 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*, pages 759–763, Granada, Spain, 1998.

[30] Melissa Holland, Jonathan Kaplan, and Michelle Sams, editors. *Intelligent Language Tutors: Theory Shaping Technology.* Lawrence Erlbaum Associates, Hillsdale, NJ, 1995.

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

[32] Sergei Nirenburg, Jaime Carbonell, Masaru Tomita, and Kenneth Goodman, editors. *Machine Translation: A Knowledge-Based Approach.* Morgan Kaufmann Publishers, San Mateo, CA, 1992.

[33] Douglas W. Oard. *Multilingual Text Filtering Techniques for High-Volume Broad-Domain Sources.* PhD thesis, University of Maryland, 1996.

[34] 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.

[35] Martha Palmer and Zhibao Wu. Verb Semantics for English-Chinese Translation. *Machine Translation*, 10(1–2):59–92, 1995.

10 RELATED PROGRAM AREAS

1. Virtual Environments.

3. Other Communication Modalities.

4. Adaptive Human Interfaces.

11 POTENTIAL RELATED PROJECTS

We are currently investigating the applicability of the linguistically-motivated framework described above to large-scale tasks in different languages and applications. One area of investigation is the use of text-based NLP techniques in man/machine interaction for a virtual visual environment. We are also investigating other communication modalities (such as speech) and human-computer interaction in applications such as foreign language tutoring and multilingual information retrieval. Finally, we are studying the application of NLP and tutoring techniques to the problem of handwriting instruction for children.