Large-Scale Interlingual Machine Translation

Bonnie J. Dorr

CONTACT INFORMATION: Department of Computer Science; A.V. Williams Building, Room 3157; University of Maryland; College Park, MD 20742; e-mail: bonnie@umiacs.umd.edu; fax: 301-314-9658; phone: 301-405-6768

WWW PAGE:
http://www.cs.umd.edu/users/bonnie/

PROGRAM AREA: Speech and NLU

KEYWORDS: principle-based machine translation; multilingual natural language processing; interlinguas; lexical acquisition; foreign language tutoring; large scale resources.

PROJECT SUMMARY: The overall objective of the NYI project (1993–1998) is to develop new computational methods, based on principles in both computer science and linguistics, for interlingual Machine Translation (MT). Specifically, the project aims to systematize the relation between syntax and semantics in lexical representations and to examine the problem of lexicon construction in MT.

Our initial focus was on the development of a principle-based analyzer for an interlingual MT system called PRINCITRAN [6, 8, 10, 12]. In general, parsers of existing principle-based systems are exceedingly inefficient since they tend to adopt the filter-based paradigm. We have combined the benefits of a message-passing paradigm with the benefits of the parameterized approach to build a more efficient, but easily extensible parsing system for MT. More recently, our research has focused on the construction of lexicons to support large-scale MT. A major result of this work has been the development of a general compositional form that serves as a language-independent meaning representation (an *interlingua*) [1, 2, 3, 4, 11, 14, 15]. We have built a large-scale lexicon in English that contains entries based on this meaning representation.

To further support our large-scale objectives, we have developed automatic techniques for lexicon construction in multiple languages. We have built a program that predicts salient linguistic features of a non-English word by using the features of its English translation in a bilingual dictionary. While not exact, owing to inexact translations and language-to-language variations, these techniques can augment an existing dictionary with reasonable accuracy, thus saving significant time. We have conducted two experiments [7]. The first one was designed primarily to demonstrate the feasibility of building a database of thematic role information for over 6500 verbs. Our approach in this experiment relies on the use of grammar codes from online syntactic dictionaries to derive the range of possible thematic grids to verbal arguments. The second experiment was designed to demonstrate that, using a combination of online syntactic dictionaries and well-known linguistic classes of verbs, it is possible to arrive at a more enriched semantic classification of at least 50% of these verbs and, hence, a more rich set of thematic grids.

In our current research, we are using a combination of semantic information in WordNet and syntactic information in LDOCE to categorize verbs in terms of a linguistically motivated semantic classification system [9]. Preliminary results indicate that it is possible to achieve an accuracy level as high as 80%. The results of this work will be used for the development of general purpose tools for automatic construction of machine-tractable dictionaries in several languages. As proof of concept, we have developed a small Spanish-English bilingual lexicon of 5500 verbs and are currently incorporating the representations from our work in English into this lexicon. The expected result of this research is a general approach to automatic construction of large-scale dictionaries for language processing applications.

PROJECT REFERENCES:

[1] Dorr, B.J., “Machine Translation Divergences: A Formal Description and Proposed Solution,” *Computational Linguistics*, 20:4, pp. 597–633, 1994.

[2] Dorr, B.J., C.R. Voss, E. Peterson, and M. Kiker, “Concept Based Lexical Selection,” Proceedings of the AAAI-94 fall symposium on Knowledge Representation for Natural Language Processing in Implemented Systems, New Orleans, LA, 1994.

[3] Dorr, B.J. and C.R. Voss, “The Case for A MT Developers’ Tool with a Two-Component View of the Interlingua,” in Proceedings of the First Annual Association for MT in the Americas Conference on Partnerships in Translation Technology, Columbia, MD, pp. 40–47, 1994.

[4] Dorr, B.J., “A Lexical-Semantic Solution to the Divergence Problem in Machine Translation,” in

Patrick Saint-Dizier and Evelyn Viegas (eds.), *Computational Lexical Semantics*, Cambridge University Press, 1994.

[5] Dorr, B.J., J. Hendler, S. Blanksteen, and B. Migdalof, "Use of LCS and Discourse for Intelligent Tutoring: On Beyond Syntax," in M. Holland, J. Kaplan, and M. Sams (eds.), *Intelligent Language Tutors: Balancing Theory and Technology*, Lawrence Erlbaum Associates, Hillsdale, NJ, 1994.

[6] Dorr, B.J., "Interlingual Machine Translation: A Parameterized Approach," in F. Pereira and B. Grosz (eds.), *Natural Language Processing*, MIT Press, Cambridge, MA, 1994.

[7] Dorr, B.J., J. Garman, and A. Weinberg, "From Syntactic Encodings to Thematic Roles: Building Lexical Entries for Interlingual MT," *Machine Translation*, 9:3, 1995.

[8] Dorr, B.J., D. Lin, J. Lee, and S. Suh, "Efficient Parsing for Korean and English: A Parameterized Message Passing Approach," in *Computational Linguistics*, 21:2, 1995.

[9] Dorr, B.J., "Building a Lexical-Semantic Dictionary for Intelligent Tutoring," submitted, *Computer-Assisted Language Learning*, 1995.

[10] Dorr, B.J., D. Lin, J. Lee, and S. Suh, "Parameterized Message-Passing for Non-head-driven Parsing," in Daniel Jones (ed.), 1995.

[11] Dorr, B.J. and C.R. Voss, "A Multi-Level Approach to Interlingual MT: Defining the Interface between Representational Languages, " *International Journal of Expert Systems*, to appear.

[12] Lin, D., Dorr, B.J., J. Lee, and S. Suh, "A Parameter-Based Message-Passing Parser for MT of Korean and English," in Proceedings of the First Annual Association for MT in the Americas Conference on Partnerships in Translation Technology, pp. 149–156, Columbia, MD, 1994.

[13] Oard, D.W., N. DeClaris, B.J. Dorr, and C. Faloutsos, "Automatic Filtering of Multilingual Texts," IEEE International Conference on Systems, Man and Cybernetics, pp. 1645–1650, San Antonio, TX, 1994.

[14] Voss, C.R. and B.J. Dorr, "Defining the Lexical Component in Interlinguas," in Proceedings of AAAI spring symposium on Representation and Acquisition of Lexical Knowledge: Polysemy, Ambiguity, and Generativity, Stanford, CA, 1995.

[15] Voss, C. and B.J. Dorr, "Identifying the Lexical Component in a Two-Component View of the Interlingua," *Machine Translation*, 10:1, to appear.

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 [17, 20] and foreign language tutoring [18]. 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 dictionaries (i.e., large NL databases that relate words to their corresponding meanings) [16]. To date, designers have been forced to build such dictionaries 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 [19]—which accommodate automatic dictionary 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.

AREA REFERENCES:

[16] Boguraev, B. and T. Briscoe, *Computational Lexicography for Natural Language Processing*, Longman, London, 1989.

[17] Dorr, B., *Machine Translation: A View from the Lexicon*, MIT Press, Cambridge, MA, 1993.

[18] Holland, M., J. Kaplan, and M. Sams, *Intelligent Language Tutors: Theory Shaping Technology*, Lawrence Erlbaum Associates, Hillsdale, NJ, 1995.

[19] Levin, B., *English Verb Classes and Alternations: A Preliminary Investigation*, Chicago, IL, 1993.

[20] Nirenburg, S., J. Carbonell, M. Tomita, and K. Goodman, *Machine Translation: A Knowledge-Based Approach* Morgan Kaufmann, San Mateo, CA, 1992.

RELATED PROGRAM AREAS: Other Communication Modalities; Adaptive Human Interfaces.

POTENTIAL RELATED PROJECTS: We are currently investigating the applicability of the linguistically-motivated framework described above to large-scale tasks in different languages and sub-areas. Future directions will involve the use of our text-based NLP techniques in systems that support other communication modalities (such as speech) and that support human-computer interaction in applications such as foreign language tutoring (FLT) [5, 9] (computer-aided instruction of a foreign language) and multilingual information filtering (IF) [13] (evaluation of articles from an information stream and selection of those which are relevant to a user's interest regardless of their source language). We plan to investigate the development of: (1) a set of novel machine-tractable dictionary representations that apply uniformly to languages as diverse as Arabic, French, and Korean; (2) human-language tutoring systems designed to use our interlingual representations; and (3) linguistically-motivated techniques for accurate selection of texts from a multilingual information stream.