System Demonstration
CatVar: A Database of Categorial Variations for English

Nizar Habash
Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20740
habash@umiacs.umd.edu

Bonnie Dorr
Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20740
bonnie@umiacs.umd.edu

Abstract

We present a new large-scale database called "CatVar" (Habash and Dorr, 2003) which contains categorial variations of English lexemes. Due to the prevalence of cross-language categorial variation in multilingual applications, our categorial-variation resource may serve as an integral part of a diverse range of natural language applications. Thus, the research reported herein overlaps heavily with that of the machine-translation, lexicon-construction, and information-retrieval communities. We demonstrate this database, embedded in a graphical interface; we also show a GUI for user input of corrections to the database.

1 Introduction

We demonstrate a new large-scale database called "CatVar" which contains *categorial variations* on a large scale for English lexemes. We also show a GUI for user input of corrections to the database. Due to the prevalence of cross-language categorial variation in multilingual applications, our categorial-variation resource may serve as an integral part of a diverse range of natural language applications. Thus, the database described herein addresses the needs of researchers in the machine-translation, lexicon-construction, and information-retrieval communities.

CatVar has already been used effectively in a wide range of monolingual and multilingual NLP applications and it is now freely available to the research community. We expect that the contribution of this resource will become more widely recognized through its future incorporation into additional NLP applications. For example, it is the intention of UMD researchers and WordNet 1.7 developers to use CatVar information for more rapid development and extension of WordNet and mutual validation of both resources.

This paper discusses other available resources and how they differ from the CatVar database. We then discuss how and what resources were used to build CatVar. For a more detailed discussion and evaluation of CatVar, see (Habash and Dorr, 2003). The CatVar is web-browseable at *http://clipdemos.umiacs.umd.edu/catvar/*.

2 Background

Lexical relations describe relative relationships among different lexemes. Lexical relations are either hierarchical taxonomic relations (such as hypernymy, hyponymy and entailments) or non-hierarchical congruence relations (such as identity, overlap, synonymy and antonymy) (Cruse, 1986). Resources specifying the relations among lexical items such as WordNet (Fellbaum, 1998) and HowNet (Dong, 2000) (among others) have inspired the work of many researchers in NLP (Carpuat et al., 2002; Dorr et al., 2000; Resnik, 1999; Hearst, 1998).

WordNet is the most well-developed and widely used lexical database of English (Fellbaum, 1998). In WordNet, both types of lexical relations are specified among words with the same part of speech (verbs, nouns, adjectives and adverbs). WordNet has been used by many researchers for different purposes ranging from the construction or extension of knowledge bases such as SENSUS (Knight and Luk, 1994) or the Lexical Conceptual Structure Verb Database (LVD) (Green et al., 2001) to the *faking* of meaning ambiguity as part of system evaluation (Bangalore and Rambow, 2000). In the context of these projects, one criticism of WordNet is its lack of cross-categorial links, such as verb-noun or noun-adjective relations.

Mel'čuk approaches lexical relations by defining a lexical combinatorial zone that specifies semantically related lexemes through Lexical Functions (LF). These functions define a correspondence between a *key* lexical item and a set of related lexical items (Mel'čuk, 1988). There are two types of functions: paradigmatic and syntagmatic (Ramos et al., 1994). Paradigmatic LFs associate a lexical item with related lexical items. The *relation* can be semantic or syntactic. Semantic LFs include Synonym(calling) = *vocation*, Antonym(small) = *big*, and Generic(fruit) = *apple*. Syntactic LFs in-

clude Derived-Noun(expand)= *expansion* and Adjective(female) = *feminine*.

Syntagmatic LFs specify collocations with a lexeme given a specified relationship. For example, there is a LF that returns a light verb associated with the LF's *key*: Light-Verb(attention) = *pay*. Other LFs specify certain semantic associations such as Intensify-Qualifier(escape) = *narrow* and Degradation(milk) = *sour*. LFs have been used in MT and Generation (e.g. (Ramos et al., 1994)).

Although research on LFs provides an intriguing theoretical discussion, there are no large scale resources available for categorial variations induced by LFs.[1] This lack of resources shouldn't suggest that the problem is too trivial to be worthy of investigation or that a solution would not be a significant contribution. On the contrary, categorial variations are necessary for handling many NLP problems. For example, in the context of MT, (Habash and Dorr, 2002) claims that 98% of all translation *divergences* (variations in how source and target languages structure meaning) involve some form of categorial variation. Moreover, most IR systems require some way to reduce variant words to common roots to improve the ability to match queries (Xu and Croft, 1998; Hull and Grefenstette, 1996; Krovetz, 1993).

Given the lack of large-scale resources containing categorial variations, researchers frequently develop and use alternative algorithmic approximations of such a resource. These approximations can be divided into Reductionist (Analytical) or Expansionist (Generative) approximations. The former focuses on the conversion of several surface forms into a common root. Stemmers such as the Porter stemmer (Porter, 1980) are a typical example. The latter, or expansionist approaches, overgenerate possibilities and rely on a statistical language model to rank/select among them. The morphological generator in Nitrogen is an example of such an approximation (Langkilde and Knight, 1998).

There are two types of problems with approximations of this type: (1) They are uni-directional and thus limited in usability—A stemmer cannot be used for generation and a morphological overgenerator cannot be used for stemming; (2) The crude approximating nature of such systems causes many problems in quality and efficiency from over-stemming/under-stemming or over-generation/under-generation.

Consider, for example, the Porter stemmer, which stems *commune*$_N$, *communication*$_N$ and *communism*$_N$ to $commun$, yet it does not produce this same stem for *communist*$_N$ or *communicable*$_{AJ}$ (stemmed to $communist$ and $communic$ respectively).[2] Another example is the expansionist Nitrogen morphological generator, where the morphological feature $+nominalize-verb$ applied to $develop$ returns eleven variations including $*developage$, $*developication$ and $*developy$. Only two are correct ($development$ and $developing$). Such overgeneration multiplied out at different points in a sentence expands the search space exponentially, and given various cutoffs in the search algorithm, might even appear in some of the top ranked choices.

These issues have served as the background for the construction of a database of categorial variations that can be used with both expansionist and reductionist approaches without the cost of over/under-stemming/generation. This database is relevant to MT, IR, and lexicon construction.

3 Building the CatVar

A categorial variation of a word with a certain part-of-speech is a derivationally-related word with possibly a different part-of-speech. For example, *hunger*$_V$, *hunger*$_N$ and *hungry*$_{AJ}$ are categorial variations of each other, as are *cross*$_V$ and *across*$_P$, and *stab*$_V$ and *stab*$_N$. Although this relation seems basic on the surface, this relation is critical to work in Information Retrieval (IR), Natural Language Generation (NLG) and Machine Translation (MT)—yet there is no large scale resource available for English that focuses on categorial variations.

The CatVar database was developed using a combination of resources and algorithms including the Lexical Conceptual Structure (LCS) Verb and Preposition Databases (Dorr, 2001), the Brown Corpus section of the Penn Treebank (Marcus et al., 1993), an English morphological analysis lexicon developed for PC-Kimmo (Englex) (Antworth, 1990), NOMLEX (Macleod et al., 1998), Longman Dictionary of Contemporary English (LDOCE)[3]

[1] The following are the only LF databases we are aware of: (1) the ETAP-3 MT system contains large Two combinatorial databases for Russian and English in the ETAP-3 MT system. These databases are on the order of 50K words, but only 2,000 entries have LFs associated with them (Boguslavsky, 1995); and (2) DiCo, a French combinatorial dictionary is underdevelopment with currently a couple of thousand entries (Polguère, 2000).

[2] For a deeper discussion and classification of Porter stemmer's errors, see (Krovetz, 1993).

[3] An English Verb-Noun list extracted from LDOCE was

(Procter, 1983), WordNet 1.6 (Fellbaum, 1998), and the Porter stemmer. The contribution of each of these sources is clearly labeled in the CatVar database, thus enabling the use of different cross-sections of the resource for different applications.[4]

Some of these resources were used to extract *seed* links between different words (Englex lexicon, NOMLEX and LDOCE). Others were used to provide a large-scale coverage of lexemes. In the case of the Brown Corpus, which doesn't provide lexemes for its words, the Englex morphological analyzer was used together with the part of speech specified in the Penn Tree Bank to extract the lexeme form. The Porter stemmer was later used as part of a clustering step to expand the seed links to create clusters of words that are categorial variants of each other, e.g., $hunger_N$, $hungry_{AJ}$, $hunger_V$, $hungriness_N$.

The current version of the CatVar (version 2.0) includes 62,232 clusters covering 96,368 unique lexemes. The lexemes belong to one of four parts-of-speech (Noun 62%, Adjective 24%, Verb 10% and Adverb 4%). Almost half of the clusters currently include one word only. Three-quarters of these single-word clusters are nouns and one-fifth are adjectives. The other half of the words is distributed in a Zipf fashion over clusters from size 2 to 27.

A smaller supplementary database devoted to verb-preposition variations was constructed solely from the LCS verb and preposition lexicon using shared LCS primitives to cluster. The database was inspired by pairs such as $cross_V$ and $across_P$ which are used in Generation-Heavy MT. But since verb-preposition clusters are not typically morphologically related, they are kept separate from the rest of the CatVar database.[5]

Figure 1 shows the CatVar web-based interface with the *hunger* cluster as an example. The interface allows searching clusters using regular expressions as well as cluster length restrictions. The database is also available for researchers in perl/C and lisp searchable formats.

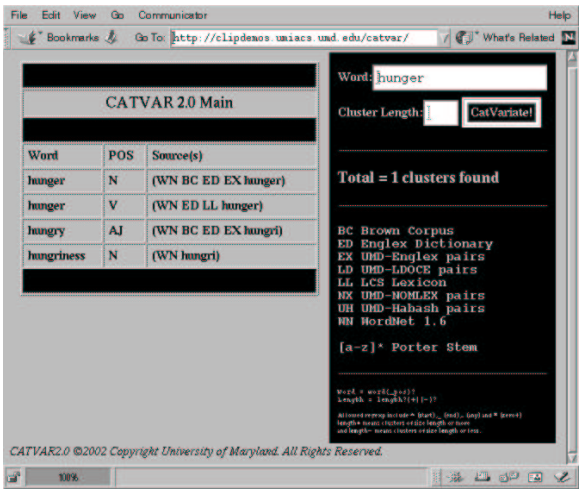

Figure 1: Web Interface

4 Applications

Our project is focused on semi-automatic resource building for MT applications. However, the CatVar database is relevant to a number of natural language applications including: (1) generation for MT, (2) headline generation, and (3) cross-language divergence unraveling for bilingual alignment. Due to space limitations, we discuss only the first of these here.[6]

The Generation-Heavy Hybrid MT (GHMT) approach accommodates asymmetrical resources for source-language (SL) poor and target-language (TL) rich languages (English, in our case). In this approach, the CatVar database is used as part of the solution to the conflation problem — cases such as the Spanish sentence *Mary le dio puñaladas a John* (literally, 'Mary gave stabs to John') being translated into *Mary stabbed John*. In GHMT, the input SL dependency structure is maintained while all words are translated to TL. Generating a conflated version of the input is conditional upon the existence of a categorial variant of a TL word that satisfies lexical semantic and thematic consistency constraints. For example, $stab_V$ is a categorial variant of $stab_N$ and it maintains $John$'s thematic role in the example above as $goal$. Details on the databases used to verify the additional constraints are available in (Habash, 2002).

5 Conclusions and Future Work

We have presented our approach to constructing a new large-scale database containing categorial vari-

provided by Rebecca Green.

[4]For example, in a headline generation system (HeadGen), higher Bleu scores were obtained when using the portions of the CatVar database that are most relevant to nominalized events (e.g., NOMLEX).

[5]This supplementary database includes 242 clusters for more than 230 verbs and 29 prepositions. Other examples of verb-preposition clusters include: $avoid_V$ and $away\ from_P$; $enter_V$ and $into_P$; and $border_V$ and $beside_P$ (or $next\ to_P$).

[6]See (Habash and Dorr, 2003) for more details about the other two applications.

ations of English words. Future work includes improving the word-cluster ratio and absorbing more of the single-word clusters into existing clusters or other single-word clusters. We are also considering enrichment of the clusters with types of derivational relations such as "nominal-event" or "doer" to complement part-of-speech labels. Other lexical semantic features such telicity, sentience and change-of-state can also be induced from morphological cues (Light, 1996).

Acknowledgments

This work has been supported, in part, by Army Research Lab Cooperative Agreement DAAD190320020, NSF CISE Research Infrastructure Award EIA0130422, and Office of Naval Research MURI Contract FCPO.810548265.

References

Antworth, E. (1990). *PC-KIMMO: A Two-Level Processor for Morphological Analysis*. Dallas Summer Institute of Linguistics.

Bangalore, S. and Rambow, O. (2000). Corpus-Based Lexical Choice in Natural Language Generation. In *Proceedings of the ACL*, Hong Kong.

Boguslavsky, I. (1995). A bi-directional Russian-to-English machine translation system (ETAP-3). In *Proceedings of the Machine Translation Summit V.*, Luxembourg.

Carpuat, M., Ngai, G., Fung, P., and Church, K. (2002). Creating a Bilingual Ontology: A Corpus-Based Approach for Aligning WordNet and HowNet. In *Proceedings of the 1st Global WordNet Conference*, Mysore, India.

Cruse, D. (1986). *Lexical Semantics*. Cambridge University Press.

Dong, Z. (2000). HowNet Chinese-English Conceptual Database. Technical Report Online Software Database, Released at ACL. http://www.keenage.com.

Dorr, B. J. (2001). LCS Verb Database. Technical Report Online Software Database, University of Maryland, College Park, MD. http://www.umiacs.umd.edu/ ˜bonnie/LCS_ Database_ Docmentation.html.

Dorr, B. J., Levow, G.-A., and Lin, D. (2000). Building a Chinese-English Mapping between Verb Concepts for Multilingual Applications. In *Proceedings of the Fourth Conference of the Association for Machine Translation in the Americas (AMTA), Cuernavaca, Mexico*, pages 1–12.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. MIT Press. http://www.cogsci.princeton.edu/˜wn [2000, September 7].

Green, R., Pearl, L., Dorr, B. J., and Resnik, P. (2001). Mapping WordNet Senses to a Lexical Database of Verbs. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 244–251, Toulouse, France.

Habash, N. (2002). Generation-Heavy Machine Translation. In *Proceedings of the International Natural Language Generation Conference (INLG'02) Student Session*, New York.

Habash, N. and Dorr, B. (2003). A Categorial Variation Database for English. In *North American Association for Computational Linguistics*, Edmonton, Canada.

Habash, N. and Dorr, B. J. (2002). Handling Translation Divergences: Combining Statistical and Symbolic Techniques in Generation-Heavy Machine Translation. In *Fifth Conference of the Association for Machine Translation in the Americas, AMTA-2002*, Tiburon, California.

Hearst, M. (1998). Automated Discovery of WordNet Relations. In Fellbaum, C., editor, *WordNet: an Electronic Lexical Database*. MIT Press.

Hull, D. A. and Grefenstette, G. (1996). Experiments in Multilingual Information Retrieval. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. http://www.xerox.fr /people /grenoble /hull /papers /sigir96.ps.

Knight, K. and Luk, S. (1994). Building a Large Knowledge Base for Machine Translation. In *Proceedings of AAAI-94*.

Krovetz, R. (1993). Viewing Morphology as an Inference Process,. In *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 191–203.

Langkilde, I. and Knight, K. (1998). Generation that Exploits Corpus-Based Statistical Knowledge. In *ACL/COLING 98, Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (joint with the 17th International Conference on Computational Linguistics)*, pages 704–710, Montreal, Canada.

Light, M. (1996). Morphological Cues for Lexical Semantics. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*.

Macleod, C., Grishman, R., Meyers, A., Barrett, L., and Reeves, R. (1998). NOMLEX: A Lexicon of Nominalizations. In *Proceedings of EURALEX'98*, Liege, Belgium.

Marcus, M., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a Large Annotated Corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Mel'čuk, I. (1988). *Dependency Syntax: Theory and Practice*. State University of New York Press, New York.

Polguère, A. (2000). Towards a theoretically-motivated general public dictionary of semantic derivations and collocations for French. In *Proceedings of EURALEX-2000*, Stuttgart.

Porter, M. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.

Procter, P. (1983). Longman Dictionary of Contemporary English: Computer Codes for the Definition Space Other than the Subject Field. Longman Group LTD.

Ramos, M. A., Tutin, A., and Lapalme, G. (1994). Lexical Functions of the Explanatory Combinatorial Dictionary for Lexicalization in Text Generation. In Saint-Dizier, P. and Viegas, E., editors, *Computational Lexical Semantics*. Cambridge University Press.

Resnik, P. (1999). Disambiguating Noun Groupings with Respect to WordNet Senses. In Armstrong, S., Church, K., Isabelle, P., Manzi, S., Tzoukermann, E., and Yarowsky, D., editors, *Natural Language Processing Using Very Large Corpora*, pages 77–98. Kluwer Academic, Dordrecht.

Xu, J. and Croft, W. B. (1998). Corpus-based stemming using cooccurrence of word variants. *ACM Transactions on Information Systems*, 16(1):61–81.