A Categorial Variation Database for English

Nizar Habash and Bonnie Dorr
Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20740
http://umiacs.umd.edu/labs/CLIP
{habash,bonnie}@umiacs.umd.edu

Abstract

We describe our approach to the construction and evaluation of a large-scale database called "CatVar" 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 apply the information-retrieval metrics of precision and recall to evaluate the accuracy and coverage of our database with respect to a human-produced gold standard. This evaluation reveals that the categorial database achieves a high degree of precision and recall. Additionally, we demonstrate that the database improves on the linkability of Porter Stemmer by over 30%.

1 Introduction

Natural Language Processing (NLP) applications may only be as good as the resources upon which they rely. Resources specifying the relations among lexical items such as WordNet (Fellbaum, 1998) and HowNet (Dong, 2000) (among others) have been used effectively in many NLP systems.

In this paper we introduce a new resource called CatVar which specifies the lexical relation *Categorial Variation* on a large scale for English. This resource has already been used effectively in a wide range of monolingual and multilingual NLP applications. Upon its first public release, Catvar will be 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.

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.[1]

In the rest of this paper we discuss other available resources and how they differ from the CatVar database. We then discuss how and what resources were used to build CatVar. We then present three applications that use CatVar in different ways: Generation-Heavy MT, headline generation, and cross-language divergence unraveling for bilingual

[1] It is the intention of the WordNet 1.7 developers to include such information in their next version, but only for nouns and verbs (Christiane Fellbaum, pc.), not other pairings such as noun-adjective, verb-preposition relationships. Discussions are currently underway for sharing the CatVar database with WordNet developers for more rapid development, extension, and mutual validation of both resources.

alignment. Finally, we present a multi-component evaluation of the database. Our evaluation reveals that the categorial database achieves a high degree of precision and recall and that it improves on the linkability of Porter Stemmer by over 30%.

2 Background

Lexical relations describe relative relationships among different lexemes. According to (Cruse, 1986), 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).

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 include 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*. Lexical Functions have been used in MT and Generation (e.g. (Ramos et al., 1994)).

Although research on Lexical Functions provides an intriguing theoretical discussion, there are no large scale resources available for categorial variations induced by lexical functions. 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 et al., 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 information retrieval 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 cause 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 $commun$. And 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 cut-offs in the search algorithm, might even appear in some of the top ranked choices.

Given these issues, our goal is to build a database of categorial variations that can be used with both expansionist and reductionist approaches without the cost of over/under-stemming/generation. The research reported herein is relevant to machine-translation, lexicon-construction, and information-retrieval.

First, we describe the construction of the "CatVar" database and its use in multilingual applications. Following this, we demonstrate the application of information-retrieval metrics of precision and recall in an evaluation of our database with respect to a human-produced gold standard. Finally, we demonstrate that the database improves on the linkability of Porter Stemmer by over 30%.

3 Building the CatVar

The CatVar database was developed using a combination of resources and algorithms including the 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] (Procter, 1983), WordNet 1.6 (Fellbaum, 1998), and the Porter stemmer (Porter, 1980). 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. Figure 1 shows the word-cluster distribution.

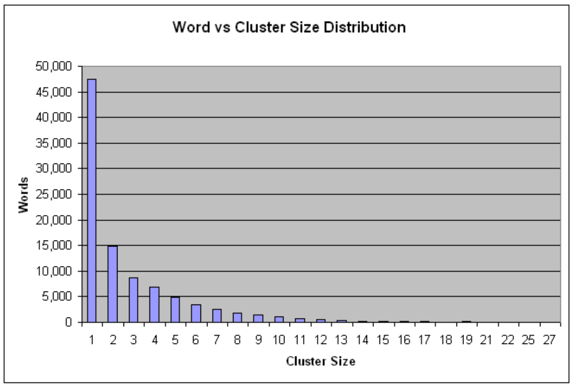

Figure 1: CatVar Distribution

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

[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 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).

the CatVar database and they were not included in the evaluation presented in this paper.[5]

The CatVar is web-browseable at *http://clipdemos.umiacs.umd.edu/catvar/*. Figure 2 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.

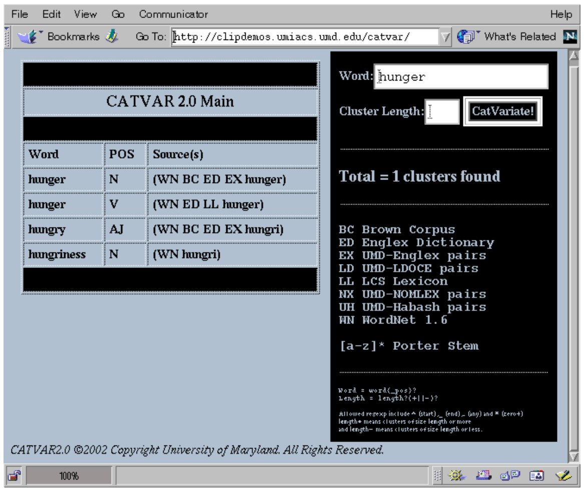

Figure 2: Web Interface

4 Applications

Our project is focused on resource building and evaluation. However, the CatVar database is relevant to a number of natural language applications, including generation for MT, headline generation, and cross-language divergence unraveling for bilingual alignment. Each of these are discussed below, in turn.

4.1 Generation-Heavy Machine Translation

The Generation-Heavy Hybrid Machine Translation (GHMT) model was introduced in (Habash, 2002) to handle translation divergences between language pairs with asymmetrical (poor-source/rich-target) resources. The approach does not rely on a transfer lexicon or a common interlingual representation to map between divergent structural configurations from source to target language. Instead, different alternative structural configurations are over-generated and these are statistically ranked using a language model.

The CatVar database is used as one of the constraints on the structural expansion step. For example, to allow the conflation of verbs such as $make_V$ or $cause_V$ and an argument such as $development_N$, the first condition for *conflatability* is finding a verb categorial variant of the argument $development_N$. In this case the verb categorial variant is $develop_V$.[6]

4.2 Headline Generation

The HeadGen headline generator was introduced in (Zajic et al., 2002) to create headlines automatically from newspaper text. The goal is to generate an *informative* headline (one that specifies the event and its participants) not just an *indicative* headline (which specifies the topic only). The system is implemented as a Hidden Markov Model enhanced with a postprocessor that filters out headlines that do not contain a verbal or nominalized event. This is achieved by verifying that there is at least one word in the generated headline that appears in CatVar as a V (a verbal event) or as a N whose verbal counterpart is in the same cluster (a nominalized event).

A recent study indicates that there is a significant improvement in Bleu scores (using human-generated headlines as our references) when running headline generation with the CatVar filter:[7]

- HeadGen with CatVar filter: 0.1740
- HeadGen with no CatVar filter: 0.1687

This quantitative distinction correlates with human-perceived differences, e.g., between the two headlines *Washingtonians fight over drugs* and *In the nation's capital* (generated for the same story—with and without CatVar, respectively).

4.3 DUSTer

DUSTer—Divergence Unraveling for Statistical Translation—was introduced in (Dorr et al., 2002).

[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]The other conditions on conflatability and some detailed examples are discussed in (Habash, 2002) and (Habash and Dorr, 2002).

[7]For details about the Bleu evaluation metric, see (Papineni et al., 2002).

In this system, common divergence types are systematically identified and English sentences are transformed to bear a closer resemblance to that of another language using a mapping referred to as E-to-E'. The objective is to enable more accurate alignment and projection of dependency trees in another language without requiring any training on dependency-tree data in that language.

The CatVar database has been incorporated into two components of the DUSTer system: (1) In the E-to-E' mapping, e.g., the transformation from $kick_V$ to *LightVB* $kick_N$ (corresponding to the English/Spanish divergence pair *kick*/*dar patada*); and (2) During an automatic mark-up phase prior to this transformation, where the particular E-to-E' mapping is selected from a set of possibilities based on the 2 input sentences. For example, the rule `V[CatVar=N] -> LightVB N` is selected for the transformation above by first checking that the verb V is associated with a word of category N in CatVar. Transforming divergent English sentences using this mechanism has been shown to facilitate word-level alignment by reducing the number of unaligned and multiply-aligned words.

5 Evaluation

This section includes two evaluations concerned with different aspects of the CatVar database. The first evaluation calculates the recall and precision of CatVar's clustering and the second determines the contribution of CatVar over Porter Stemmer.

5.1 CatVar Clustering Evaluation: Recall and Precision

To determine the recall and precision of CatVar given the lack of a gold standard, we asked 8 native speakers to evaluate 400 randomly-selected clusters. Each annotator was given a set of 100 clusters (with two annotators per set). Figure 3 shows a segment of the evaluation interface which was web-browseable.

The annotators were given detailed instructions and many examples to help them with the task. They were asked to classify each word in every cluster as belonging to one of the following categories:

- Perfect: This word definitely belongs in this cluster.
- Perfect (except for part of speech problem).

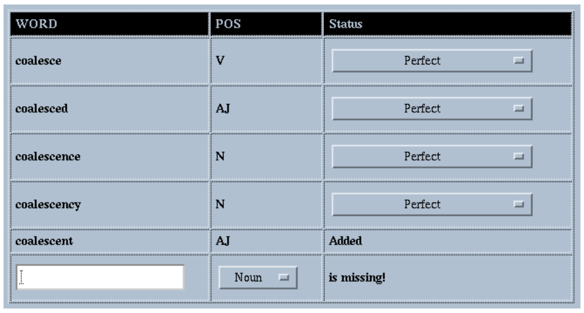

Figure 3: Evaluation

- Perfect (except for spelling problem).
- Not Sure: It is not clear whether a word that is derivationally correct belongs in a set or not.
- Doesn't Belong: This word doesn't belong in this cluster.
- May not be a Real Word: This word is not known and couldn't be found it in a dictionary.

The interface also provided an input text box to add missing words to a cluster.

In calculating the inter-annotator agreement, we did not consider mismatches in word additions as disagreement since some annotators could not think up as many possible variations as others. After all, this was not an evaluation of their ability to think up variations, but rather of the coverage of the CatVar database. Even though there were six fine-grained classifications, the average inter-annotator agreement was high (80.75%). Many of the disagreements, however, resulted from the fine-grainness of the options available to the annotators.

In a second calculation of inter-annotator agreement, we simplified the annotators' choices by placing them into three groups corresponding to Perfect (Perfect and Perfect-*but*), Not-sure (Not-sure and May-not-be-a-real-word) and Wrong (Does-not-belong). This annotation-grouping approach is comparable to the clustering techniques used by (Veronis, 1998) to "super-tag" fine grained annotations. After grouping the annotations, average inter-annotator agreement rose up to 98.35%.

The cluster modifications produced by each pair of annotators assigned to the same cluster were then combined automatically in an approximation to post-annotation inter-annotator discussion, which traditionally results in agreement: (1) If both annotators agreed on a category, then it stands; (2) One annotator overrides another in cases where one is more sure than the other (i.e., Perfect overrides Perfect-but-with-error/Not-sure and Wrong overrides Not-sure); (3) In cases where one annotator considers a word Perfect while the other annotator considered it Wrong, we compromise at Not-sure. The union of all added words was included in the combined cluster.

The 400 combined clusters covered 808 words. 68% of the words were ranked as Perfect. None had spelling errors and only one word had a part-of-speech issue. 23 words (less than 3%) were marked as Not-sures. And only 6 words (less than 1%) were marked as Wrong. There were 209 added words (about 26%). However 128 words (or 61% of missing words) were not actually missing, but rather not linked into the set of clusters evaluated by a particular annotator. Some of these words were clustered separately in the database.[8] The rest of the missing words (81 words or 10% of all words) were not present in the database, but 50 of them (or 62%) were linkable to existing words in the CatVar using simple stemming (e.g., the Porter stemmer, whose relevance is described next).

The precision was calculated as the ratio of perfect words to all original (i.e. not added) words: 91.82%. The recall was calculated as the ratio of perfect words divided by all perfect plus all added words: 72.46%. However, if we exclude the not-really missing words, the adjusted recall value becomes 87.16%. The harmonic mean or F-score[9] of the precision and recall is 81.00% (or 89.43% for adjusted recall).

5.2 Linkability Evaluation: Comparison to Porter Stemmer

To measure the contribution of Catvar with respect to the "linking together" of related words, it is important to define the concept of *linkability* as the percentage of word-to-word links in the database resulting from a specific source. For example, *Natural linkability* refers to pairs of words whose form doesn't change across categories such as zip_V and zip_N or $afghan_N$ and $afghan_{AJ}$. *Porter linkability* refers to words linkable by reduction to a common Porter stem. *CatVar linkability* is the linkability of two words appearing in the same CatVar cluster.

Figure 4 shows an example of all three types of links in the *hunger* cluster. Here, $hunger_N$ and $hunger_V$ are linked in three ways, Naturally (N), by the Porter stemmer (P), and in CatVar (C). Porter links $hungry_{AJ}$ and $hungriness_N$ via the common stem *hungri* but Porter doesn't link either of these to $hunger_N$ or $hunger_V$ (stem *hunger*). The total number of links in this cluster is six, two of which are Porter-determinable and only one of which is naturally-determinable.

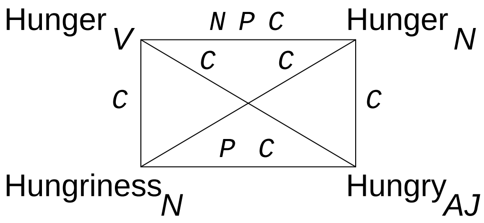

Figure 4: Three Types of Links

The calculation of linkability applies only to the portion of the database containing multi-word clusters (about half of the database) since single-word clusters have zero links. The 48,867 linked words are distributed over 14,731 clusters with 89,638 total number of links. About 12% of these links are naturally-determinable and 70% are Porter-linkable. The last 30% of the links is a significant contribution of the CatVar database, compared to the Porter Stemmer, particularly since this stemmer is an industry standard in the Information Retrieval community.

It is important to point out that, for CatVar to be used in IR, it must be accompanied by an inflectional analyzer that reduces words to their lexeme form (removing plural endings from nouns or gerund ending from verbs).[10] The contribution of CatVar is in its linking of words related derivationally not inflectionally. Work by (Krovetz, 1993) demonstrates an improved performance with derivational stemming over the Porter Stemmer most of the time.

[8] The 128 words that were "not really missing" were clustered in 89 other clusters not included in the evaluation sample.

[9] F-score = $\frac{2 \times Precision \times Recall}{Precision + Recall}$.

[10] This is, in fact, the approach used in the HeadGen and DUSTer applications described above.

6 Conclusions and Future Work

We have presented our approach to constructing and evaluating a new large-scale database containing categorial variations of English words. In addition, we have described different applications for which it has proven useful. Our evaluation indicates that CatVar has coverage and accuracy of over 80% (F-score) and also that the database improves the linkability of Porter stemmer by about 30%. These findings are significant contributions to several different communities, including information retrieval and machine translation.

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 enriching the clusters with types of derivational relations such as "nominal-event" or "doer" to complement part-of-speech labels. Additionally, we are interested in measuring the applied contribution of using the CatVar in natural-language applications. And finally, we intend to incorporate CatVar into new applications such as parallel corpus word alignment.

Acknowledgments

This work has been supported, in part, by ONR MURI Contract FCPO.810548265, Mitre Contract 010418-7712, and NSF CISE Research Infrastructure Award EIA0130422. We would like to thank all the annotators who participated in the evaluation of the database.

References

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

S. Bangalore and O. Rambow. 2000. Exploiting a probabilistic hierarchical model for generation.

D. Cruse. 1986. *Lexical Semantics*. Cambridge University Press.

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

Bonnie J. Dorr, Lisa Pearl, Rebecca Hwa, and Nizar Habash. 2002. DUSTer: A Method for Unraveling Cross-Language Divergences for Statistical Word-Level Alignment. In *Proceedings of the Fifth Conference of the Association for Machine Translation in the Americas, AMTA-2002*, Tiburon, California.

Bonnie J. Dorr. 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

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

Rebecca Green, Lisa Pearl, Bonnie J. Dorr, and Philip Resnik. 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.

Nizar Habash and Bonnie J. Dorr. 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.

Nizar Habash, Bonnie J. Dorr, and David Traum. 2002. Efficient Language Independent Generation from Lexical Conceptual Structures. *Machine Translation*.

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

David A. Hull and Gregory Grefenstette. 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.

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

R. Krovetz. 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.

Irene Langkilde and Kevin Knight. 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.

Catherine Macleod, Ralph Grishman, Adam Meyers, Leslie Barrett, and Ruth Reeves. 1998. NOMLEX: A Lexicon of Nominalizations. In *Proceedings of EURALEX'98*, Liege, Belgium.

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

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

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of Association of Computational Linguistics*, Philadelphia, PA.

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

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

Margarita Alonso Ramos, Agnes Tutin, and Guy Lapalme. 1994. Lexical Functions of the Explanatory Combinatorial Dictionary for Lexicalization in Text Generation. In Patrick Saint-Dizier and Evelyne Viegas, editors, *Computational Lexical Semantics*. Cambridge University Press.

J. Veronis. 1998. A study of polysemy judgements and inter-annotator agreement. In *Programme and advanced papers of the Senseval workshop*, Herstmonceux Castle, England.

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

David M. Zajic, Bonnie J. Dorr, and Rich Schwartz. 2002. Automatic headline generation for newspaper stories. In *Proceedings of the ACL-2002 Workshop on Text Summarization*, Philadelphia, PA.