Desiderata for Tagging with WordNet Synsets or MCCA Categories

Kenneth C. Litkowski
CL Research
20239 Lea Pond Place
Gaithersburg, MD 20879
(Email: ken@clres.com)
(Web site: http://www.clres.com)

1 Abstract

Minnesota Contextual Content Analysis (MCCA) is a technique for characterizing the concepts and themes occurring in text (sentences, paragraphs, interview transcripts, books). MCCA tags each word with a category and examines the distribution of categories against norms representing general usage of categories. MCCA also scores texts in terms of social contexts that are similar to different functions of language. Distributions can be analyzed using non-agglomerative clustering to characterize the concepts and themes. MCCA categories have been mapped to WordNet senses. The defining characteristics that emerge from the mapping and the statistical techniques used in MCCA for analyzing concepts and themes suggest that tagging with WordNet synsets or MCCA categories may produce epiphenomenal results that are misleading. We suggest that WordNet synsets and MCCA categories be augmented with further lexical semantic information for use after text is tagged or categorized. We suggest that such information is useful not only for the primary purposes of disambiguation in parsing and text classification in content analysis and information retrieval, but also for tasks in corpus analysis, discourse analysis, and automatic text summarization.

2 Introduction

Content analysis provides distributional methods for analyzing characteristics of textual material. Its roots are the same as computational linguistics (CL), but it has been largely ignored in CL until recently (Dunning, 1993; Carletta, 1996; Kilgarriff, 1996). One content analysis approach, Minnesota Contextual Content Analysis (MCCA) (McTavish & Pirro, 1990), in use for over 20 years and with a well-developed dictionary category system, contains analysis methods that provide insights into the use of WordNet (Miller, et al., 1990) for tagging.

We describe the unique characteristics of MCCA, how its categories relate to WordNet synsets, the analysis methods used in MCCA to provide quantitative information about texts, what implications this has for the use of WordNet in tagging, and how these techniques may contribute to lexical semantic tagging. Specifically, we show that WordNet provides a backbone, but that additional lexical semantic information needs to be associated with WordNet synsets. We describe novel perspectives on how this information can be used in various NLP tasks.

3 Minnesota Contextual Content Analysis

MCCA differs from other content analysis techniques in using a norm for examining the distribution of its categories in a given text. The 116 categories used in the dictionary to characterize words,[1] like other content analysis category systems, are heuristic in nature. Each category has a name (e.g., *activity, fellow feeling, about changing, human roles, expression arena*).

The distinguishing characteristic of MCCA is that the emphasis of each category is normed in two ways. Categories that are emphasized in a text (E-scores) are normed against expected general usage of categories based on the Brown corpus (Kucera & Francis, 1967). The second way is based on relative usage of categories expected in four broad institutional areas. The latter is based on some initial research and subsequent work which essentially factor-analyzed profiles of category usage for texts representing a broad range of organizations and social situations (Cleveland, et al., 1974). These are referred to as context scores (C-scores) and labelled *traditional* (judicial and religious texts), *practical* (business texts), *emotional* (leisure, recreational, and fictional texts), and *analytic*

---

[1] A word may have more than one category and is disambiguated in tagging.

(scientific writings). These contexts correspond well to the functions of language (Nida, 1975: 201-5).

After tagging a text and determining category frequencies, the C-scores are calculated by comparison with the expected distribution of the contexts and the E-scores are calculated by comparison with the expected distribution of each category.[2] These are the quantitative bases for analysis of the concepts and themes.

Unlike other techniques for determining which words are characteristic of a text (Kilgarriff, 1996), such as the χ^2-test and mutual information, the C-scores and E-scores are examined not only for differences among texts, but also for over- and under-emphasis against the norms. This provides greater sensitivity to the analysis of concepts and themes.

4 MCCA Categories andWordNet Synsets

(McTavish, et al., 1995) and (McTavish, et al., 1997) suggest that MCCA categories recapitulate WordNet synsets. We used WordNet synsets in examining MCCA categories to determine their coherence, to characterize their relations with WordNet, and to understand the significance of these relations in the MCCA analysis of concepts and themes and in tagging with WordNet synsets.

In the MCCA dictionary of 11,000 words,[3] the average number of words in a category is 95, with a range from 1 to about 300. Using the DIMAP software (CL Research, 1997 - in preparation),[4] we created sublexicons of individual categories, extracted WordNet synsets for these sublexicons, extracted information from the Merriam-Webster Concise Electronic Dictionary integrated with DIMAP, and attached lexical semantic information from other resources to entries in these sublexicons.

We began with the hypothesis that the categories correspond to those developed by (Hearst & Schütze, 1996) in creating categories from the WordNet noun hierarchy. We found that the MCCA categories were generally internally consistent, but with characteristics not intuitively obvious.[5] As a result, we needed to articulate firm principles for characterizing the categories.

Eleven categories (such as *Have, Prepositions, You, I-Me, He, A-An, The*) consist of only a few words from closed classes. The category *The* contains one word with an average expected frequency of 6 percent (with a range over the four contexts of 5.5 to 6.5). The category *Prepositions* contains 18 words with an average expected frequency of 11.1 percent (with a range over the four contexts of 9.5 to 12.3 percent). About 20 categories (*Implication, If, Colors, Object, Being*) consist of a relatively small number of words (34, 22, 65, 11, 12, respectively) taken primarily from syntactically or semantically closed-class words (subordinating conjunctions, relativizers, the tops of WordNet, colors).

The remaining 80 or so categories consist primarily of open-class words (nouns, verbs, adjectives, and adverbs), sprinkled with closed-class words (auxiliaries, subordinating conjunctions). These categories require more detailed analyses.[6]

Several categories correspond well to the Hearst & Schütze model. The categories *Functional roles, Detached roles,* and *Human roles* align with subtrees rooted at particular nodes in the WordNet hierarchies. For example, *Detached roles* has a total of 66 words, with an average expected frequency of .16 percent and a range from .10 to .35 percent. The .35 percent frequency is for the *analytic* context; each of the other three contexts have expected frequencies of about .10 percent. The words in this category include:

> ACADEMIC, ARTIST, BIOLOGIST, CREATOR, CRITIC, HISTORIAN, INSTRUCTOR, OBSERVER, PHILOSOPHER,

---

[2] Disambiguation is based on a running context score. Each category has a frequency of occurrence in a context. The category selected for an ambiguous word is the one with the smallest difference from the running context score.

[3] This dictionary has tagged 85 to 95 percent of the words in about 1500 analyses covering 45 million words over the last 15 years.

[4] A suite of programs for creating and maintaining lexicons for natural language processing, available from CL Research. Procedures used in this paper, applicable to any category analysis using DIMAP, are available at http://www.clres.com. The general principles of category development followed in these procedures are described in (Litkowski, in preparation).

[5] In general, we have found that assignment of only about 5 to 10 percent of the words in a category is questionable.

[6] Analysis of MCCA categories is a continuing process.

PHYSICIST, PROFESSOR, RESEARCHER, REVIEWER, SCIENTIST, SOCIOLOGIST

These words are a subset of the WordNet synsets headed at PERSON, in particular, synsets headed by

CREATOR;
EXPERT: AUTHORITY: PROFESSIONAL;
INTELLECTUAL.[7]

Other synsets under EXPERT and AUTHORITY do not fall into this category. Thus, the heuristic *Detached roles* is like a Hearst & Schütze super-category, but not constructed on a statistical metric, rather on underlying semantic components.

Other categories do not fall out so neatly. The category *Sanction* (120 words) has an average expected frequency of .08 percent, with a range over the four contexts of .06 to .10 percent. It includes the following words (and their inflected forms):

> APPLAUD, APPLAUSE, APPROVE, CONGRATULATE, CONGRATULATION, CONVICT, CONVICTION, DISAPPROVAL, DISAPPROVE, HONOR, JUDGE, JUDGMENT, JUDGMENTAL, MERIT, MISTREAT, REJECT, REJECTION, RIDICULE, SANCTION, SCORN, SCORNFUL, SHAME, SHAMEFULLY

Examination of the WordNet synsets is similarly successful here, identifying many words (particularly verbs) in a subtree rooted at JUDGE. However, the set is defined as well by including a derivational lexical rule to allow forms in other parts of speech. Another meaning component is seen in APPROVE and DISAPPROVE, namely, the negative or pejorative prefix, again requiring a lexical rule as part of the category's definition. Such lexical rules would be encoded as described in (Copestake & Briscoe, 1991). This set of words (rooted primarily in the verbs of the set) corresponds to the (Levin, 1993) *Characterize* (class 29.2), *Declare* (29.4), *Admire* (31.2), and Judgment verbs (33) and hence may have particular syntactic and semantic patterning. The verb frames attached to WordNet verb synsets are not sufficiently detailed to cover the granularity necessary to characterize an MCCA category. Instead, the definition of this class might, following (Davis, 1996), inherit a sort *notion-rel*, which has a "perceiver" and a "perceived" argument (thus capturing syntactic patterning) with perhaps a selectional restriction on the "perceiver" that the type of action is an evaluative one (thus providing semantic patterning).

Another complex category is *Normative*, consisting of 76 words, with an average expected frequency of .60 percent and a range over the four contexts of .37 to .79 percent. This category also has words from all parts of speech and thus will entail the use of derivational lexical rules in its definition. This category includes the following (along with various inflectional forms):

> ABSOLUTE, ABSOLUTELY, CONSEQUENCE, CONSEQUENTLY, CORRECT, CORRECTLY, DOGMATISM, HABITUAL, HABITUALLY, IDEOLOGICALLY, IDEOLOGY, NECESSARILY, NECESSARY, NORM, OBVIOUSLY, PROMINENT, PROMINENTLY, REGULARITY, REGULARLY, UNEQUIVOCALLY, UNUSUAL, UNUSUALLY

The use of the heuristic *Normative* to label this category clearly reflects the presence in these words of a semantic component oriented around characterizing something in terms of expectations. But, of particular interest here, are the adverb forms. McTavish has also used the heuristic *Reasoning* for this category. These adverbs are *content disjuncts* (Quirk, et al., 1985: 8.127-33), that is, words betokening a speaker's comment on the content of what the speaker is saying, in this case, compared to some norm or standard. Thus, part of the defining characteristics for this category is a specification for lexical items that have a [content-disjunct +] feature.

These examples of words in the *Sanction* and *Normative* categories (repeated in other categories) indicates a need to define categories not only in terms of supercategories using the Hearst & Schütze model, but also with additional lexical semantic information not present in WordNet or MCCA categories. In particular, we see the need for encoding derivational and morphological relations, finer-grained characterization of government patterns, feature specifications, and primitive semantic components.

In any event, we have seen that MCCA categories are consistent with WordNet synsets. They recapitulate the WordNet synsets by acting as supercategories similar to those identified in Hearst & Schütze. To this extent, results from MCCA tagging would be similar to those of Hearst & Schütze. The MCCA methods suggest further insights based on what purposes we are trying to achieve from tagging.

---

[7] Identification of these synsets facilitates extension of the MCCA dictionary to include further hyponyms of these synsets.

5 Analysis of Tagged Texts

The important questions at this point are why there is value in having additional lexical semantic information associated with tagging and why MCCA categories and WordNet synsets are insufficient. The answer to these questions begins to emerge by considering the further analysis performed after a text has been "classified" on the basis of the MCCA tagging. As described above, MCCA produces a set of C-scores and E-scores for each text. These scores are then subjected to analysis to provide additional results useful in social science and information retrieval applications.

The two sets of scores are used for computing the distance among texts. This distance is used directly or in exploration of the differences between texts. Unlike other content analysis techniques (or classification techniques used for measuring the distance between documents in information retrieval), MCCA uses the non-agglomerative technique of multidimensional scaling (MDS).[8] This technique (Kruskal & Wish, 1977) produces a map when given a matrix of distances.

MDS does not presume that a 2-dimensional representation displays the distances between texts. Rather, it unfolds the dimensions one-by-one, starting with 2, examines statistically how "stressed" the solution is, and then adds further dimensions until the stress shows signs of reaching an asymptote. Output from the scaling provides "rotation" maps at each dimension projected onto 2-dimensional space.

McTavish, et al. illustrates the simple and the more complex use of these distance metrics. In the simple use, the distance between transcripts of nursing home patients, staff, and administrators was used as a measure of social distance among these three groups. This measure was combined with various characteristics of nursing homes (size, type, location, etc.) for further analysis, using standard statistical techniques such as correlation and discriminant analysis.

In the more complex use, the MDS results identify the concepts and themes that are different and similar in the transcripts. This is accomplished by visually inspecting the MDS graphical output. Examination of the 4-dimensional context vectors provides an initial characterization of the texts. The analyst identifies the contextual focus (*traditional, practical, emotional,* or *analytic*) and the ways in which the texts differ from one another. This provides general themes and pointers for identifying the conceptual differences among the texts.

MDS analysis of the E-score vectors identifies the major concepts that differentiate the texts. The analyst examines the graphical output to label points with the dominant MCCA categories. The "meaning" (that is, the underlying concepts) of the MDS graph is then described in terms of category and word emphases. These are the results an investigator uses in reporting on the content analysis using MCCA.

This is the point at which the insufficiency of MCCA categories (and WordNet synsets) becomes visible. In examining the MDS output, the analysis is subjective and based only on identification of particular sets of words that distinguish the concepts in each text (much like the techniques described in (Kilgarriff, 1996) that are used in authorship attribution). If the MCCA categories had richer definitions based on additional lexical semantic information, the analysis could be performed based on less subjective and more rigorously defined principles.

(Burstein, et al., 1996) describe techniques for using lexical semantics to classify responses to test questions. An essential component of this classification process is the identification of sublexicons that cut across parts of speech, along with concept grammars based on collapsing phrasal and constituent nodes into a generalized XP representation. As seen above in the procedures for defining MCCA categories, addition of lexical semantic information in the form of derivational and morphological relations and semantic components common across part of speech boundaries—information now lacking in WordNet synsets—would facilitate the development of concept grammars.

(Briscoe & Carroll, 1997) describe novel techniques for constructing a subcategorization dictionary from analysis of corpora. They note that their system needs further refinement, suggesting that adding information to lexical entries about diathesis alternation possibilities and semantic selectional preferences on argument heads is likely to improve their results. Again, the procedures for analyzing MCCA categories seem to require this type of information.

We have discussed elsewhere (Litkowski & Harris, 1997) extension of a discourse analysis algorithm

---

[8] Agglomerative techniques cluster the two closest texts (with whatever distance metric) and then successively add texts one-by-one as they are closest to the existing cluster.

incorporating lexical cohesion principles. In this extension, we found it necessary to require use of the AGENTIVE and CONSTITUTIVE qualia of nouns (see (Pustejovsky, 1995: 76)) as selectional specifications on verbs to maintain lexical cohesion. With such information, we were able not only to provide a more coherent discourse analysis of a text segment, but also possibly to summarize the text better.

6 Discussion and Future Work

We have shown how MCCA categories generally recapitulate WordNet synsets and how MCCA analysis leads to thematic and conceptual characterization of texts. Since MCCA categories do not exactly correspond to WordNet subtrees, but frequently represent a bundle of syntactic and semantic properties, we believe that the tagging results are epiphenomenal. Since the MCCA results seem more robust than tagging with WordNet synsets (q.v. (Voorhees, 1994)), we suggest that this is due to more specific meaning components underlying the MCCA categories.

(Nida, 1975: 174) characterized a semantic domain as consisting of words sharing semantic components. However, he also suggests (Nida, 1975: 193) that domains represent an arbitrary grouping of the underlying semantic features. We suggest that the MCCA categories and WordNet synsets represent two such systems of domains, each reflecting particular perspectives.

This suggests that categorical systems used for tagging need to be augmented with more precise lexical semantic information. This information can be semantic features, semantic roles, subcategorization patterns, syntactic alternations (e.g., see (Dorr, in press)), and semantic components. We suggest that the use of this lexical semantic information in tagging may provide considerable benefit in analyzing tagging results.

We are continuing analysis of the MCCA categories to characterize them in terms of lexical semantic information. We are using a variety of lexical resources, including WordNet, the database by (Dorr, in press) based on (Levin, 1993), and COMLEX (Macleod & Grishman, 1994; Wolff, et al., 1995). We will propagate these meaning components to the lexical items.

After automating the MDS analysis, we will examine the extent to which the lexical semantic information is correlated with the thematic analyses. We hypothesize that the additional information will provide greater sensitivity for characterizing the concepts and themes.

7 Acknowledgments

I would like to thank Don McTavish, Thomas Pötter, Robert Amsler, Mary Dee Harris, some WordNet folks (George Miller, Shari Landes, and Randee Tengi), Tony Davis, and anonymous reviewers for their discussions and comments on issues relating to this paper and its initial draft.

8 References

Briscoe, T., & Carroll, J. (1997). Automatic extraction of subcategorization from corpora. 5th Conference on Applied Natural Language Processing. Washington, DC: Association for Computational Linguistics.

Burstein, J., Kaplan, R., Wolff, S., & Lu, C. (1996, June). Using lexical semantic information techniques to classify free responses. In E. Viegas & M. Palmer (Eds.), *Breadth and Depth of Semantic Lexicons*. Workshop Sponsored by the Special Interest Group on the Lexicon. Santa Cruz, CA: Association for Computational Linguistics.

Carletta, J. (1996). Assessing agreement on classification tasks: The Kappa statistic. *Computational Linguistics*, *22*(2), 249-254.

CL Research. (1997 - in preparation). *DIMAP-3 users manual*. Gaithersburg, MD.

Cleveland, C. E., McTavish, D. G., & Pirro, E. B. (1974, September 5-13). Contextual content analysis. ISSC/CISS Workshop on Content Analysis In the Social Sciences. Pisa, Italy: Standing Committee on Social Science Data of the International Social Science Council, UNESCO, Centrol Nazionale Universitario de Calcolo Electtronico (CUNCE).

Copestake, A. A., & Briscoe, E. J. (1991, June 17). Lexical operations in a unification-based framework. ACL SIGLEX Workshop on Lexical Semantics and Knowledge Representation. Berkeley, CA: Association for Computational Linguistics.

Davis, A. R. (1996). Lexical semantics and linking in the hierarchical lexicon [diss], Stanford, CA: Stanford University.

Dorr, B. (in press). Large-scale dictionary construction for foreign language tutoring and interlingual machine translation. *Journal of Machine Translation*.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, *19*(1), 61-74.

Hearst, M. A., & Schütze, H. (1996). Customizing a lexicon to better suit a computational task. In B. Boguraev & J. Pustejovsky (Eds.), *Corpus processing for lexical acquisition* (pp. 77-96). Cambridge, MA: The MIT Press.

Kilgarriff, A. (1996, April). Which words are particularly characteristic of a text? A survey of statistical approaches. European Conference on Artificial Intelligence.

Kruskal, J. B., & Wish, M. (1977). *Multidimensional scaling*. Beverly Hills, CA: Sage Publications.

Kucera, H., & Francis, W. N. (1967). *Computerized dictionary of present-day American English*. Providence, RI: Brown University Press.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. Chicago, IL: The University of Chicago Press.

Litkowski, K. C. (in preparation). Category development based on semantic principles. *Social Science Computer Review*.

Litkowski, K. C., & Harris, M. D. (1997). *Category development using complete semantic networks*. Gaithersburg, MD: CL Research.

Macleod, C., & Grishman, R. (1994). *COMLEX syntax reference manual*. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

McTavish, D. G., Litkowski, K. C., & Schrader, S. (1995, September). A computer content analysis approach to measuring social distance in residential organizations for older people. Society for Content Analysis by Computer. Mannheim, Germany.

McTavish, D. G., Litkowski, K. C., & Schrader, S. (1997). A computer content analysis approach to measuring social distance in residential organizations for older people. *Social Science Computer Review*, in press.

McTavish, D. G., & Pirro, E. B. (1990). Contextual content analysis. *Quality & Quantity, 24*, 245-265.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography, 3*(4), 235-244.

Nida, E. A. (1975). *Componential analysis of meaning*. The Hague: Mouton.

Pustejovsky, J. (1995). *The generative lexicon*. Cambridge, MA: The MIT Press.

Quirk, R., Greenbaum, S., Leech, G., & Svartik, J. (1985). *A comprehensive grammar of the English language*. London: Longman.

Voorhees, E. M. (1994, July 3-6). Query expansion using lexical-semantic relations. In W. B. Croft & C. J. van Rijsbergen (Eds.), *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval* (pp. 61-69). Dublin, Ireland: Springer-Verlag.

Wolff, S. R., Macleod, C., & Meyers, A. (1995). *COMLEX word classes*. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.