CONCEPTUAL ASSOCIATION FOR COMPOUND NOUN ANALYSIS

Mark Lauer

Microsoft Institute
65 Epping Road
North Ryde NSW 2113
(t-markl@microsoft.com)

AUSTRALIA

Department of Computing
Macquarie University
NSW 2109
(mark@macadam.mpce.mq.edu.au)

Abstract

This paper describes research toward the automatic interpretation of compound nouns using corpus statistics. An initial study aimed at syntactic disambiguation is presented. The approach presented bases associations upon thesaurus categories. Association data is gathered from unambiguous cases extracted from a corpus and is then applied to the analysis of ambiguous compound nouns. While the work presented is still in progress, a first attempt to syntactically analyse a test set of 244 examples shows 75% correctness. Future work is aimed at improving this accuracy and extending the technique to assign semantic role information, thus producing a complete interpretation.

INTRODUCTION

Compound Nouns: Compound nouns (CNs) are a commonly occurring construction in language consisting of a sequence of nouns, acting as a noun; *pottery coffee mug,* for example. For a detailed linguistic theory of compound noun syntax and semantics, see Levi (1978). Compound nouns are analysed syntactically by means of the rule N → N N applied recursively. Compounds of more than two nouns are ambiguous in syntactic structure. A necessary part of producing an interpretation of a CN is an analysis of the attachments within the compound. Syntactic parsers cannot choose an appropriate analysis, because attachments are not syntactically governed. The current work presents a system for automatically deriving a syntactic analysis of arbitrary CNs in English using corpus statistics.

Task description: The initial task can be formulated as choosing the most probable binary bracketing for a given noun sequence, known to form a compound noun, without knowledge of the context. E.G.: (*pottery* (*coffee mug*)); ((*coffee mug*) *holder*)

Corpus Statistics: The need for wide ranging lexical-semantic knowledge to support NLP, commonly referred to as the ACQUISITION PROBLEM, has generated a great deal of research investigating automatic means of acquiring such knowledge. Much work has employed carefully constructed parsing systems to extract knowledge from machine readable dictionaries (e.g., Vanderwende, 1993). Other approaches have used rather simpler, statistical analyses of large corpora, as is done in this work.

Hindle and Rooth (1993) used a rough parser to extract lexical preferences for prepositional phrase (PP) attachment. The system counted occurrences of unambiguously attached PPs and used these to define LEXICAL ASSOCIATION between prepositions and the nouns and verbs they modified. This association data was then used to choose an appropriate attachment for ambiguous cases. The counting of unambiguous cases in order to make inferences about ambiguous ones is adopted in the current work. An explicit assumption is made that lexical preferences are relatively independent of the presence of syntactic ambiguity.

Subsequently, Hindle and Rooth's work has been extended by Resnik and Hearst (1993). Resnik and Hearst attempted to include information about typical prepositional objects in their association data. They introduced the notion of CONCEPTUAL ASSOCIATION in which associations are measured between groups of words considered to represent concepts, in contrast to single words. Such class-based approaches are used because they allow each observation to be generalized thus reducing the amount of data required. In the current work, a freely available version of Roget's thesaurus is used to provide the grouping of words into concepts, which then form the basis of conceptual association. The research presented here can thus be seen as investigating the application of several key ideas in Hindle and Rooth (1993) and in Resnik and Hearst (1993) to the solution of an analogous problem, that of compound noun analysis. However, both these works were aimed solely at syntactic disambiguation. The goal of semantic interpretation remains to be investigated.

METHOD

Extraction Process: The corpus used to collect information about compound nouns consists of some 7.8 million words from Grolier's multimedia on-line encyclopedia. The University of Pennsylvania morphological analyser provides a database of more than 315,000 inflected forms and their parts of speech. The Grolier's text was searched for consecutive words listed in the database as always being nouns and separated only by white space. This prevented comma-separated lists and other non-compound noun sequences from being included. However, it did eliminate many CNs from consideration because many nouns are occasionally used as verbs and are thus ambiguous for part of speech. This resulted in 35,974 noun sequences of which all but 655 were pairs. The first 1000 of the sequences were examined manually to check that they were not incidentally adjacent nouns (as in direct and indirect objects, say). Only 2% did not form CNs, thus establishing a reasonable utility for the extraction method. The pairs were then used as a training set, on the assumption that a two word noun compound is unambiguously bracketed.[1]

Thesaurus Categories: The 1911 version of Roget's Thesaurus contains 1043 categories, with an average of 34 single word nouns in each. These categories were used to define concepts in the sense of Resnik and Hearst (1993). Each noun in the training set was tagged with a list of the categories in which it appeared.[2] All sequences containing nouns not listed in Roget's were discarded from the training set.

Gathering Associations: The remaining 24,285 pairs of category lists were then processed to find a conceptual association (CA) between every ordered pair of thesaurus categories ((t_1, t_2)) using the formula below. $CA(t_1, t_2)$ is the mutual information between the categories, weighted for ambiguity. It measures the degree to which the modifying category predicts the modified category and vice versa. When categories predict one another, we expect them to be attached in the syntactic analysis.

Let **AMBIG**(w) = the number of thesaurus
categories w appears in (the ambiguity of w).

Let **COUNT**(w_1, w_2) = the number of instances of
w_1 modifying w_2 in the training set

Let **FREQ**(t_1, t_2) =

$$\sum_{w1 \text{ in } t1} \sum_{w2 \text{ in } t2} \frac{\mathrm{COUNT}(w_1, w_2)}{\mathrm{AMBIG}(w_1) \cdot \mathrm{AMBIG}(w_2)}$$

Let **CA** (t_1, t_2) =

$$\frac{\mathrm{FREQ}(t_1, t_2)}{\sum_{\forall i} \mathrm{FREQ}(t_1, i) \cdot \sum_{\forall i} \mathrm{FREQ}(i, t_2)}$$

where *i* ranges over all possible thesaurus categories. Note that this measure is asymmetric. $CA(t_1, t_2)$ measures the tendency for t_1 to modify t_2 in a compound noun, which is distinct from $CA(t_2, t_1)$.

Automatic Compound Noun Analysis: The following procedure can be used to syntactically analyse ambiguous CNs. Suppose the compound consists of three nouns: $w_1 w_2 w_3$. A left-branching analysis, $[[w_1 w_2] w_3]$ indicates that w_1 modifies w_2, while a right-branching analysis, $[w_1 [w_2 w_3]]$ indicates that w_1 modifies something denoted primarily by w_3. A modifier should be associated with words it modifies. So, when CA(*pottery*, *mug*) >> CA(*pottery*, *coffee*), we prefer (*pottery* (*coffee mug*)). First though, we must choose concepts for the words. For each w_i ($i = 2$ or 3), choose categories S_i (with w_1 in S_i) and T_i (with w_i in T_i) so that $CA(S_i, T_i)$ is greatest. These categories represent the most significant possible word meanings for each possible attachment. Then choose w_i so that $CA(S_i, T_i)$ is maximum and bracket w_1 as a sibling of w_i. We have then chosen the attachment having the most significant association in terms of mutual information between thesaurus categories.

In compounds longer than three nouns, this procedure can be generalised by selecting, from all possible bracketings, that for which the product of greatest conceptual associations is maximized.

RESULTS

Test Set and Evaluation: Of the noun sequences extracted from Grolier's, 655 were more than two nouns in length and were thus ambiguous. Of these, 308 consisted only of nouns in Roget's and these formed the test set. All of them were triples. Using the full context of each sequence in the test set, the author analysed each of these, assigning one of four possible outcomes. Some sequences were not CNs (as observed above for the extraction process) and were labeled Error. Other sequences exhibited what Hindle and Rooth (1993) call SEMANTIC INDETERMINACY, where the meanings associated with two attachments cannot be distinguished in the context. For example, *college economics texts*. These were labeled Indeterminate. The remainder were labeled Left or Right depending on whether the actual analysis is left- or right-branching.

TABLE 1 - Test set analysis distribution:

Labels	L	R	I	E	Total
Count	163	81	35	29	308
Percentage	53%	26%	11%	9%	100%

Proportion of different labels in the test set.

Table 1 shows the distribution of labels in the test set. Hereafter only those triples that received a bracketing (Left or Right) will be considered.

The attachment procedure was then used to automatically assign an analysis to each sequence in

[1] This introduces some additional noise, since extraction can not guarantee to produce complete noun compounds

[2] Some simple morphological rules were used at this point to reduce plural nouns to singular forms

the test set. The resulting correctness is shown in Table 2. The overall correctness is 75% on 244 examples. The results show more success with left branching attachments, so it may be possible to get better overall accuracy by introducing a bias.

TABLE 2 - Results of test:

x	Output Left	Output Right
Actual Left	131	32
Actual Right	30	51

The proportions of correct and incorrect analyses.

DISCUSSION

Related Work: There are two notable systems that are related to the current work. The SENS system described in Vanderwende (1993) extracted semantic features from machine readable dictionaries by means of structural patterns applied to definitions. These features were then matched by heuristics which assigned likelihood estimates to each possible semantic relationship. The work only addressed the interpretation of pairs of nouns and did not mention the problem of syntactic ambiguity.

A very simple technique aimed at bracketing ambiguous compound nouns is reported in Pustejovsky et al. (1993). While attempting to extract taxonomic relationships, their system heuristically bracketed CNs by searching elsewhere in the corpus for subcomponents of the compound. Such matching fails to take account of the natural frequency of the words and is likely to require a much larger corpus for accurate results. Unfortunately, they provide no evaluation of the performance afforded by their approach.

Future Plans: A more sophisticated noun sequence extraction method should improve the results, providing more and cleaner training data. Also, many sequences had to be discarded because they contained nouns not in the 1911 Roget's. A more comprehensive and consistent thesaurus needs to be used.

An investigation of different association schemes is also planned. There are various statistical measures other than mutual information, which have been shown to be more effective in some studies. Association measures can also be devised that allow evidence from several categories to be combined.

Compound noun analyses often depend on contextual factors. Any analysis based solely on the static semantics of the nouns in the compound cannot account for these effects. To establish an achievable performance target for context free analysis, an experiment is planned using human subjects, who will be given ambiguous noun compounds and asked to choose attachments for them.

Finally, syntactic bracketing is only the first step in interpreting compound nouns. Once an attachment is established, a semantic role needs to be selected as is done in SENS. Given the promising results achieved for syntactic preferences, it seems likely that semantic preferences can also be extracted from corpora. This is the main area of ongoing research within the project.

CONCLUSION

The current work uses thesaurus category associations gathered from an on-line encyclopedia to make analyses of compound nouns. An initial study of the syntactic disambiguation of 244 compound nouns has shown promising results, with an accuracy of 75%. Several enhancements are planned along with an experiment on human subjects to establish a performance target for systems based on static semantic analyses. The extension to semantic interpretation of compounds is the next step and represents promising unexplored territory for corpus statistics.

ACKNOWLEDGMENTS

Thanks are due to Robert Dale, Vance Gledhill, Karen Jensen, Mike Johnson and the anonymous reviewers for valuable advice. This work has been supported by an Australian Postgraduate Award and the Microsoft Institute, Sydney.

REFERENCES

Hindle, Don and Mats Rooth (1993) "Structural Ambiguity and Lexical Relations" *Computational Linguistics* Vol. 19(1), Special Issue on Using Large Corpora I, pp 103-20

Levi, Judith (1978) "The Syntax and Semantics of Complex Nominals" Academic Press, New York.

Pustejovsky, James, Sabine Bergler and Peter Anick (1993) "Lexical Semantic Techniques for Corpus Analysis" *Computational Linguistics* Vol. 19(2), Special Issue on Using Large Corpora II, pp 331-58

Resnik, Philip and Marti Hearst (1993) "Structural Ambiguity and Conceptual Relations" *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*, June 22, Ohio State University, pp 58-64

Vanderwende, Lucy (1993) "SENS: The System for Evaluating Noun Sequences" in Jensen, Karen, George Heidorn and Stephen Richardson (eds) "Natural Language Processing: The PLNLP Approach", Kluwer Academic, pp 161-73