RAISINS, SULTANAS, AND CURRANTS: LEXICAL CLASSIFICATION AND ABSTRACTION VIA CONTEXT PRIMING

David J. Hutches
Department of Computer Science and Engineering, Mail Code 0114
University of California, San Diego
La Jolla, CA 92093-0114
dhutches@ucsd.edu

Abstract

In this paper we discuss the results of experiments which use a context, essentially an ordered set of lexical items, as the seed from which to build a network representing statistically important relationships among lexical items in some corpus. A metric is then applied to the nodes in the network in order to discover those pairs of items related by high indices of similarity. The goal of this research is to instantiate a class of items corresponding to each item in the priming context. We believe that this instantiation process is ultimately a special case of abstraction over the entire network; in this abstraction, similar nodes are collapsed into meta-nodes which may then function as if they were single lexical items.

I. Motivation and Background

With respect to the processing of language, one of the tasks at which human beings seem relatively adept is the ability to determine when it is appropriate to make generalizations and when it is appropriate to preserve distinctions. The process of abstraction and knowing when it might reasonably be used is a necessary tool in reducing the complexity of the task of processing natural language. Part of our current research is an investigation into how the process of abstraction might be realized using relatively low-level statistical information extracted from large textual corpora.

Our experiments are an attempt to discover a method by which class information about the members of some sequence of lexical items may be obtained using strictly statistical methods. For our purposes, the class to which a lexical item belongs is defined by its instantiation. Given some context such as `he walked across the room`, we would like to be able to instantiate classes of items corresponding to each item in the context (e.g., the class associated with `walked` might include items such as `paced`, `stepped`, or `sauntered`).

The corpora used in our experiments are the Lancaster-Oslo-Bergen (LOB) corpus and a subset of the ACL/DCI Wall Street Journal (WSJ) corpus. The LOB corpus consists of a total of 1,008,035 words, composed of 49,174 unique words. The subset of the WSJ corpus that we use has been pre-processed such that all letters are folded to lower case, and numbers have been collapsed to a single token; the subset consists of 18,188,548 total words and 159,713 unique words.

II. Context Priming

It is not an uncommon notion that a word may be defined not rigourously as by the assignment of static syntactic and semantic classes, but dynamically as a function of its usage (Firth 1957, 11). Such usage may be derived from co-occurrence information over the course of a large body of text. For each unique lexical item in a corpus, there exists an "association neighbourhood" in which that item lives; such a neighbourhood is the probability distribution of the words with which the item has co-occurred. If one posits that similar lexical items will have similar neighbourhoods, one possible method of instantiating a class of lexical items would be to examine all unique items in a corpus and find those whose neighbourhoods are most similar to the neighbourhood of the item whose class is being instantiated. However, the potential computational problems of such an approach are clear. In the context of our approach to this problem, most lexical items in the search space are not even remotely similar to the item for which a class is being instantiated. Furthermore, a substantial part of a lexical item's association neighbourhood provides only superficial information about that item. What is required is a process whereby the search space is reduced dramatically. One method of accomplishing this pruning is via context priming.

In context priming, we view a context as the seed upon which to build a network describing that part of the corpus which is, in some sense, close to the context. Thus, just as an individual lexical item has associated with it a unique neighbourhood, so too does a context have such a neighbourhood. The basic process of building a network is straightforward. Each item in the priming context has associated with it a unique neighbourhood defined in terms of those lexical items with which it has co-occurred. Similarly, each of these latter items also has a unique association neighbourhood. Generating a network based on some context consists in simply expanding nodes (lexical items) further and further away from the context until some threshold, called the depth of the network, is reached.

Just as we prune the total set of unique lexical items by context priming, we also prune the neighbourhood of each node in the network by using a statistical metric which provides some indication of how important the relationship is between each lexical item and the items in its neighbourhood. In the results we describe here, we use mutual information (Fano 1961, 27–28; Church and Hanks 1990) as the metric for neighbourhood pruning, pruning which occurs as the network is being generated. Yet, another parameter controlling the topology of the network is the extent of the "window" which defines the neighbourhood of a lexical item (e.g., does the neighbourhood of a lexical item consist of only those items which have co-occurred at a distance of up to 3, 5, 10, or 1000 words from the item).

III. Operations on the Network

The network primed by a context consists merely of those lexical items which are closely reachable via co-occurrence from the priming context. Nodes in the network are lexical items; arcs represent co-occurrence relations and carry the value of the statistical metric mentioned above and the distance of co-occurrence. With such a network we attempt to approximate the statistically relevant neighbourhood in which a particular context might be found.

In the tests performed on the network thus far we use the similarity metric

$$S(x, y) = \frac{|A \cap B|^2}{|A \cup B|}$$

where x and y are two nodes representing lexical items, the neighbourhoods of which are expressed as the sets of arcs A and B respectively. The metric S is thus defined in terms of the cardinalities of sets of arcs. Two arcs are said to be equal if they reference (point to) the same lexical item at the same offset distance. Our metric is a modification of the Tanimoto coefficient (Bensch and Savitch 1992); the numerator is squared in order to assign a higher index of similarity to those nodes which have a higher percentage of arcs in common.

Our first set of tests concentrated directly on items in the seed context. Using the metric above, we attempted to instantiate classes of lexical items for each item in the context. In those cases where there were matches, the results were often encouraging. For example, in the LOB corpus, using the seed context `John walked across the room`, a network depth of 6, a mutual information threshold of 6.0 for neighbourhood pruning, and a window of 5, for the item `John`, we instantiated the class {`Edward, David, Charles, Thomas`}. A similar test on the WSJ corpus yielded the following class for `john`

$$\left\{\begin{array}{c}\texttt{robert,william,james,charles,}\\ \texttt{richard,paul,thomas,edward,david,}\\ \texttt{donald,daniel,frank,michael,dennis,}\\ \texttt{joseph,jim,alan,dan,roger}\end{array}\right\}$$

Recall that the subset of the WSJ corpus we use has had all items folded to lower case as part of the pre-processing phase, thus all items in an instantiated class will also be folded to lower case.

In other tests, the instantiated classes were less satisfying, such as the following class generated for `wife` using the parameters above, the LOB, and the context `his wife walked across the room`

$$\left\{\begin{array}{c}\texttt{mouth,father,uncle,lordship,}\\ \texttt{fingers,mother,husband,father's,}\\ \texttt{shoulder,mother's,brother}\end{array}\right\}$$

In still other cases, a class could not be instantiated at all, typically for items whose neighbourhoods were too small to provide meaningful matching information.

IV. Abstraction

It is clear that even the most perfectly derived lexical classes will have members in common. The different senses of `bank` are often given as the classic example of a lexically ambiguous word. From our own data, we observed this problem because of our preprocessing of the WSJ corpus; the instantiation of the class associated with `mark` included some proper names, but also included items such as `marks`, `currencies`, `yen`, and `dollar`, a confounding of class information that would not have occurred had not case folding taken place. Ideally, it would be useful if a context could be made to exert a more constraining influence during the course of instantiating classes. For example, if it is reasonably clear from a context, such as `mark loves mary`, that the "mark" in question is the human rather than the financial variety, how may we ensure that the context provides the proper constraining information if `loves` has never co-occurred with `mark` in the original corpus?

In the case of the ambiguous mark above, while this item does not appear in the neighbourhood of loves, other lexical items do (e.g., everyone, who, him, mr), items which may be members of a class associated with mark. What is proposed, then, is to construct incrementally classes of items over the network, such that these classes may then function as a single item for the purpose of deriving indices of similarity. In this way, we would not be looking for a specific match between mark and loves, but rather a match among items in the same class as mark; items in the same class as loves, and items in the same class as mary. With this in mind, our second set of experiments concentrated not specifically on items in the priming context, but on the entire network, searching for candidate items to be collapsed into meta-nodes representing classes of items.

Our initial experiments in the generation of pairs of items which could be collapsed into meta-nodes were more successful than the tests based on items in the priming context. Using the LOB corpus, the same parameters as before, and the priming context John walked across the room, the following set of pairs represents some of the good matches over the generated network.

$$\left\{\begin{array}{c}
\texttt{(15,20),(dont't,didn't),(3,4),(her,his),} \\
\texttt{(minutes,days),(three,five),(few, five),} \\
\texttt{(2,3),(fig,table),(days,years),(40,50),} \\
\texttt{(me,him),(three,few),(4,5),(50,100),} \\
\texttt{(currants,sultanas),(sultanas,raisins),} \\
\texttt{(currants,raisins),}\ldots
\end{array}\right\}$$

Using the WSJ corpus, again the same parameters, and the context john walked across the room, part of the set of good matches generated was

$$\left\{\begin{array}{c}
\texttt{(months,weeks),(rose,fell),(days,weeks),} \\
\texttt{(single-a-plus,triple-b-plus),} \\
\texttt{(single-a-minus,triple-b-plus),} \\
\texttt{(lawsuit,complaint),(analyst,economist)} \\
\texttt{(john,robert),(next,past),(six,five),} \\
\texttt{(lower,higher),(goodyear,firestone),} \\
\texttt{(profit,loss),(billion,million),} \\
\texttt{(june,march),(concedes,acknowledges),} \\
\texttt{(days,weeks),(months,years),}\ldots
\end{array}\right\}$$

It should be noted that the sets given above represent the best good matches. Empirically, we found that a value of $S > 1.0$ tends to produce the most meaningful pairings. At $S \leq 1.0$, the amount of "noisy" pairings increases dramatically. This is not an absolute threshold, however, as apparently unacceptable pairings do occur at $S > 1.0$, such as, for example, the pairs (catching, teamed), (accumulating, rebuffed), and (father, mind).

V. Future Research

The results of our initial experiments in generating classes of lexical items are encouraging, though not conclusive. We believe that by incrementally collapsing pairs of very similar items into meta-nodes, we may accomplish a kind of abstraction over the network which will ultimately allow the more accurate instantiation of classes for the priming context. The notion of incrementally merging classes of lexical items is intuitively satisfying and is explored in detail in (Brown, et al. 1992). The approach taken in the cited work is somewhat different than ours and while our method is no less computationally complex than that of Brown, et al., we believe that it is somewhat more manageable because of the pruning effect provided by context priming. On the other hand, unlike the work described by Brown, et al., we as yet have no clear criterion for stopping the merging process, save an arbitrary threshold. Finally, it should be noted that our goal is not, strictly speaking, to generate classes over an entire vocabulary, but only that portion of the vocabulary relevant for a particular context. It is hoped that, by priming with a context, we may be able to effect some manner of word sense disambiguation in those cases where the meaning of a potentially ambiguous item may be resolved by hints in the context.

VI. References

Bensch, Peter A. and Walter J. Savitch. 1992. "An Occurrence-Based Model of Word Categorization". *Third Meeting on Mathematics of Language.* Austin, Texas: Association for Computational Linguistics, Special Interest Group on the Mathematics of Language.

Brown, Peter F., et al. 1992. "Class-Based *n*-gram Models of Natural Language". *Computational Linguistics* 18.4: 467–479.

Church, Kenneth Ward, and Patrick Hanks. 1990. "Word Association Norms, Mutual Information, and Lexicography". *Computational Linguistics* 16.1: 22–29.

Fano, Robert M. 1961. *Transmission of Information: A Statistical Theory of Communications.* New York: MIT Press.

Firth, J[ohn] R[upert]. 1957. "A Synopsis of Linguistic Theory, 1930-55." *Studies in Linguistic Analysis.* Philological Society, London. Oxford, England: Basil Blackwell. 1–32.