

# Multiway clustering for creating biomedical term sets

Vasileios Kandylas, Lyle Ungar and Ted Sandler  
Computer and Information Science  
University of Pennsylvania  
Philadelphia, PA 19104

kandylas, ungar, tsandler@cis.upenn.edu

Shane Jensen  
Department of Statistics  
University of Pennsylvania  
Philadelphia, PA 19104

stjensen@wharton.upenn.edu

## Abstract

We present an EM-based clustering method that can be used for constructing or augmenting ontologies such as MeSH. Our algorithm simultaneously clusters verbs and nouns using both verb-noun and noun-noun co-occurrence pairs. This strategy provides greater coverage of words than using either set of pairs alone, since not all words appear in both datasets. We demonstrate it on data extracted from Medline and evaluate the results using MeSH and Wordnet.

## 1. Introduction

There has been extensive work in automating the process of generating term-sets, both in general natural language processing [8] and specifically for biomedical and bioinformatic uses [2, 10]. These methods typically cluster words under the “distributional similarity” assumption that words that occur in the same contexts are semantically related [6]. In this paper, we present a method for word clustering that combines information from different types of relations in which words occur (e.g., co-occurrence with verbs, and with other nouns) and achieves greater coverage of words and qualitatively different clusters. The clusters we obtain are broader than the usual MeSH categories and may be more useful for tasks such as relationship extraction.

Our contribution is a mixture model-based approach for finding word clusters by integrating information from different data sources. Specifically, we concentrate on integrating the co-occurrence information of word pairs extracted using different patterns, such as noun-noun and verb-noun co-occurrences. Since not all words appear in every type of pattern, taking advantage of side information helps both to improve (or as discussed below, at least alter) the quality of the clusters and also to expand the coverage of clustered words. Additionally, the different sources of information introduce new relationships between terms and the

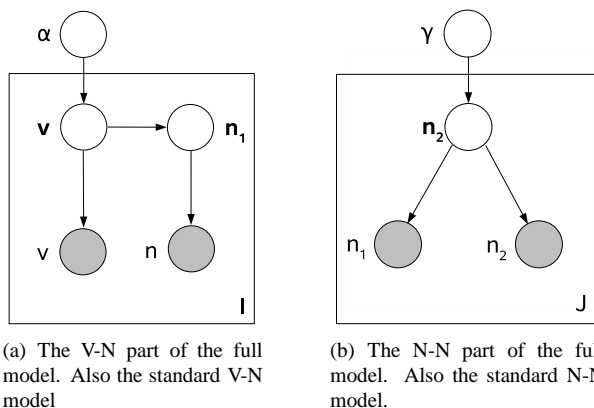


Figure 1. The full model.

resulting clusters capture concepts that are different from those found only from one type or co-occurrence pattern.

Our algorithm is based on a combined hierarchical model consisting of two simple models, similar to belief nets (BNs). In order to make simultaneous use of the two tables of data that are available, the two models are fused, by forcing them to have the same parameters for certain conditional probabilities, according to the problem-specific representation. The difference from other multi-clustering algorithms is that our proposed method is based on likelihood maximization of a hierarchical model and is thus simple, principled and quite fast. In contrast, the method of Bekkerman et.al. [1] maximizes the mutual information between pairs of variables by using a user-defined clustering schedule to perform agglomerative and divisive clustering for different subsets of the variables. The multivariate information bottleneck [5] is more similar to our method, as it is based on Bayesian networks, but it relies on maximizing mutual information under constraints defined by a second network. Both networks have to be specified by the user.

## 2. Method

Our model is designed for data that are in the form of instances of co-occurrence pairs. The dataset can be viewed as two sparse tables, the first having verbs corresponding to rows and nouns to columns and the second having nouns as both rows and columns. An element  $(i, j)$  of either table is the number of times  $n_{ij}$  that the words corresponding to row  $i$  and column  $j$  co-occur. The data tables contain  $I$  instances of verb-noun pairs and  $J$  instances of noun-noun pairs.

The idea behind our method is to use standard generative models to encode the information derived from the dataset. The number and structure of these models depends on the number and type of data tables that are available. Each data table corresponds to a submodel, whose structure is defined by the actual variables that appear in the dataset. For example, in our case, the first data table contains co-occurrence counts of a verb and its direct object, so we created a hierarchical submodel with a dependence of the noun cluster on the verb cluster (figure 1(a)). Our second data table contains co-occurrences of nouns in appositions, conjunctions and disjunctions, so we put a single noun cluster generating both nouns in the pair (figure 1(b)). Once the submodels have been chosen, the parameters that correspond to the same variables in the data tables are tied, thus achieving the combining of the information contained in the data tables. For our model, the probabilities of generating a noun given a noun cluster are constrained to be the same in both submodels. Our “full model” consists of the two separate submodels described previously, one for the verb-noun pairs, which we call the “standard V-N model”, and another for the noun-noun pairs, which we call the “standard N-N model”.

The random variables we use are defined as follows:  $V$  takes values from the set of verb clusters,  $\mathbf{N}_1$  and  $\mathbf{N}_2$  take values from the set of noun clusters,  $V$  from the set of verbs and  $N, N_1, N_2$  from the set of nouns. We use bold font for variables associated with clusters and regular font for variables associated with observations. The parameter representing the conditional probability of a noun given a noun cluster is the same in both parts of the model, thus achieving the desired coupling.

The model is estimated using the EM algorithm. The variables  $V, N, N_1, N_2$  are observed and  $\mathbf{V}, \mathbf{N}_1, \mathbf{N}_2$  are treated as unobserved. The expectation and maximization steps are computed as usual. The only noteworthy difference is for the expected number of times  $E[\#\mathbf{n}_{1n}|o]$  that the variables  $\mathbf{N}_1$  and  $N$  take some specific values  $\mathbf{n}_{1k}$  and  $n_k$ . This computation involves summing over both datasets:

$$E[\#\mathbf{n}_{1n}|o] = \sum_{i=1}^I P(\mathbf{N}_1 = \mathbf{n}_{1k}, N = n_k | v_i, n_i) +$$

$$+ \sum_{j=1}^J P(\mathbf{N}_2 = \mathbf{n}_{2k}, N = n_k | n_{1j}, n_{2j}).$$

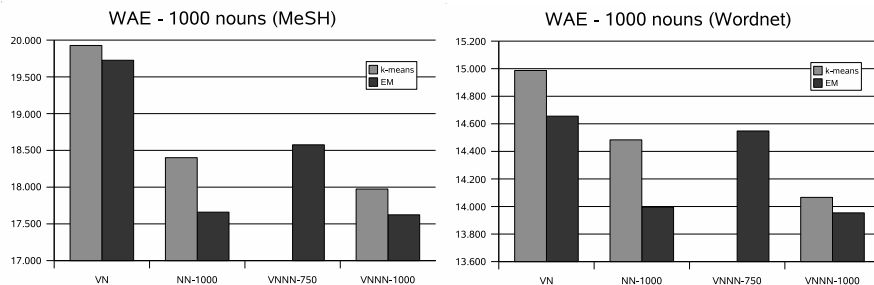
## 3. Experimental setup

We generated our dataset (“MEDLINE”) by parsing 1,800,547 abstracts from the MEDLINE database, ranging from years 1995 to 2000, with the MINIPAR parser [7]. MINIPAR can be configured to output a sequence of “dependency triples” that represent shallow syntactic configurations between words. A dependency triple has the form  $(w_1, rel_r, w_2)$  where  $w_1$  and  $w_2$  are words in a sentence that engage in some syntactic relation  $rel_r$ . We extracted two types of relations. The first was verb-direct object, giving us the set of verb-noun pairs. The second was noun-TYPE-noun, where TYPE could be apposition, conjunction or disjunction, giving the set of noun-noun pairs. From the extracted verbs and nouns we selected the 1000 most common verbs and 1000 most common nouns as our vocabulary and randomly chose 500,000 pairs of verbs and nouns and an equal number of noun-noun pairs. To simulate the situation of unequal noun coverage in a systematic and easily quantifiable way, we also constructed a reduced noun-noun table by randomly removing 250 of the 1000 nouns as well as all the noun-noun co-occurrence instances where one or both nouns were in the removed subset of 250 nouns. We thus artificially created a reduced noun-noun dataset, which does not cover all the nouns and used it to test how extra information about the missing 250 nouns (in the form of the verb-noun table) can help.

The clusters were evaluated using class labels. We labeled the verbs and nouns by mapping them to WORDNET [3] and using the hypernyms of the synsets they mapped to as labels. The nouns were also mapped to MeSH [9] and labeled by their grandparent node. The mappings gave us multiple labels per word for most words. Each cluster was represented by a distribution of labels, found by creating a histogram of label occurrences for the (labeled) words in the cluster. These label distributions were in turn used by the evaluation measure, which was the weighted average entropy (WAE), defined as the average of the label distribution entropies for the clusters weighted by the cluster sizes [4]. The lower the value of WAE, the better the clustering matches the given labels of the data.

## 4. Results

We performed a number of different experiments and used  $k$ -means as a standard measure of comparison. In all experiments we used 25 verb clusters and 50 noun clusters, as we found these numbers to give a reasonable tradeoff between cluster coherence and diversity. The EM algorithm



**Figure 2. Weighted average entropy of the noun clusters for 1000 nouns.**

was initialized with the output of  $k$ -means applied to the verb-noun data table only. Because EM gives soft clusters, we performed a hardening procedure at the end of the clustering, so that every word was assigned to one and only one cluster.

For the first set of experiments we compared the noun clusters found by EM and using:

1. only the verb-noun data and the V-N model (results denoted as VN),
2. only the complete noun-noun data and the N-N model (denoted as NN-1000),
3. the noun-noun data for 750 nouns and augmented with side information about all 1000 nouns in the form of the verb-noun table and using the full model (denoted as VNNN-750),
4. the complete noun-noun data for all 1000 nouns augmented with the verb-noun data and the full model (denoted as VNNN-1000).

For comparison with  $k$ -means we used information as similar as possible. For case 1 above, we used the verb-noun pairs, treating the verbs as features. For case 2, we used the noun-noun pairs and clustered the nouns using nouns as features. For case 4, the verb-noun and noun-noun tables were combined by concatenating them and  $k$ -means found noun clusters using both verb and noun co-occurrences as features for the nouns. We did not compare  $k$ -means with VNNN-750 of case 3, because combining the two data tables as before would give vectors with missing features and using  $k$ -means with missing features would require special treatment.

The results of these experiments are shown in figure 2. We now examine two scenarios. In the first one, the noun-noun pairs provide information for all 1000 nouns. In this case, the VN results (using only verb-noun data) perform worst, while the NN and the combined VNNN-1000 results were better in terms of WAE. For the second scenario, we assume we only have noun-noun information for 750 nouns.

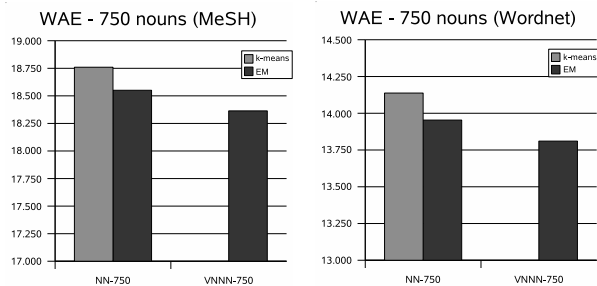
Without any other information, the only way to cluster more nouns (in this experiment the extra 250 ones) is to assign them randomly to clusters. The WAE in this case is much worse (statistically significant). The alternative is to use the verb-noun data. By combining the two tables we get the VNNN-750 results which are better than the VN ones, even though they do not give information for the 250 nouns. We can thus achieve coverage of 1,000 nouns with our clusters with better quality than if we had simply used the verb-noun data, which in this scenario are the only available data covering 1,000 nouns.

In table 1 we give representative examples of clusters from these methods. The clusters were manually selected to correspond to the same high-level notions. Both the NN-1000 and VNNN-1000 clusters in that table look of good quality, better than VN-1000, which contains several spurious words. Comparing the NN-1000 and VNNN-1000 clusters in the same table, we note that they capture different types of concepts. The VNNN-1000 cluster is about *biological substrates*, whereas NN-1000 found two separate clusters, one for *animals* and one for *micro-organisms*. The way the noun-noun pairs were extracted from the text produces highly-correlated co-occurrences that in turn yield highly specific clusters. On the other hand, the nouns that appear in verb-noun pairs are considered similar if similar actions are done to them (i.e. similar verbs apply to similar direct objects). The use of the verb-noun pairs in the full model introduces new relations between words that are not present in the noun-noun data. This causes the found clusters to be broader.

For the second set of experiments we tested how well the 750 randomly chosen nouns were clustered. We therefore ignored the 250 missing nouns when computing the WAE. Specifically, we compared the clusters found by: VNNN-750 and the N-N model, operating on the reduced noun-noun data (NN-750). The results show that for both mappings,  $k$ -means is again worse than EM (figure 3). The full model for this case performs better than the N-N model in terms of WAE. The verb-noun data containing information

VN	mice, type, tumor, tissue, line, material, strain, mutant, liver, carcinoma, culture, virus, human, specie, muscle, some, heart, clone, variant, bacteria, nucleus, derivative, tumour, cortex, spleen, fibroblast, fiber, organ, vessel, nuclei, mouse, artery, neutrophil, neck, platelet, nerve, embryo, particle, monocyte, chromosome, lymph node, organism, adenocarcinoma, adenoma, axon, intestine, astrocyte, leukocyte, CD4
NN-1000	cell, line, neuron, source, macrophage, majority, lymphocyte, clone, fibroblast, neutrophil, platelet, epithelial cell, monocyte, precursor, subset, astrocyte, leukocyte, CD4
NN-1000	model, rat, mice, animal, normal, human, dog, mouse, plant, embryo, rabbit, cat, pig
VNNN-1000	cell, rat, mice, line, that, strain, animal, mutant, neuron, normal, human, specie, macrophage, majority, lymphocyte, clone, bacteria, vector, dog, fibroblast, mouse, neutrophil, platelet, plant, epithelial cell, embryo, monocyte, precursor, organism, rabbit, cat, wild-type, subset, pathogen, pig, astrocyte, leukocyte, CD4

**Table 1. Examples of corresponding noun clusters found with three of the methods.**



**Figure 3. Evaluation of the noun clusters for the 750 randomly selected nouns.**

for all 1000 nouns help improve the clusters for the 750 nouns, making a greater contribution than in the previous set of experiments.

## 5. Discussion

Using EM on BNs with tied parameters is a useful method for incorporating information from several different sources to achieve greater coverage and cluster quality. One advantage we found was the potential to increase the coverage of clustered words. By using different syntactic patterns one can easily gather information about extra words from a relatively small corpus of text and combine it with the proposed method, acquiring clusters of a large number of words. So, nouns which do not show up in any noun-noun collocations can still be clustered if verb-noun pairs are available. The alternative would be to use a single syntactic pattern on a larger corpus, which may not always be available.

The second “advantage” is more subtle. Supplementing noun-noun co-occurrence data with verb-noun data changes the nature of the clusters that are found. Nouns that are the targets of the same action (and thus appear with the same verb) often constitute a different, and broader, set than nouns that are mentioned together only in noun-noun collo-

cations. For example, different types of tissues and of animal were clustered together when the verb-noun information was used, since they are all similar experimental substrates. Finding clusters of different type can be helpful for building ontologies that organize information in alternative ways, for example for use in extracting relationships between entities, where broader classes would be useful.

## References

- [1] R. Bekkerman, R. El-Yaniv, and A. McCallum. Multi-way distributional clustering via pairwise interactions. *Proceedings of ICML*, 22:41–48, 2005.
- [2] J. Fan and C. Friedman. Semantic Classification of Biomedical Concepts Using Distributional Similarity. *Journal of the American Medical Informatics Association*, 14(4):467–477, 2007.
- [3] C. Fellbaum. *Wordnet: An Electronic Lexical Database*. MIT Press, 1998.
- [4] X. Z. Fern and C. E. Brodley. Random projection for high dimensional data clustering: A cluster ensemble approach. In *Proceedings of the Twentieth International Conference of Machine Learning*, pages 186–193, Washington, DC, USA, 2003. AAAI Press.
- [5] N. Friedman, O. Mosenzon, N. Slonim, and N. Tishby. Multivariate information bottleneck. *Uncertainty in Artificial Intelligence: Proceedings of the Seventeenth Conference (UAI-2001)*, pages 152–161, 2001.
- [6] H. Li. Word clustering and disambiguation based on co-occurrence data. *Natural Language Engineering*, 8(01):25–42, 2002.
- [7] D. Lin. Dependency-based evaluation of MINIPAR. In *Proceedings of the Workshop on the Evaluation of Parsing Systems, at LREC-98*. Springer, 1998.
- [8] D. Lin and P. Pantel. Induction of semantic classes from natural language text. *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 317–322, 2001.
- [9] National Library of Medicine. Medical subject headings – annotated alphabetic list, 2002. Bethesda, MD, The Library, 2001.
- [10] J. Weeds, J. Dowdall, G. Schneider, B. Keller, and D. J. Weir. Using distributional similarity to organise biomedical terminology. *Terminology*, 11:107–141(35), 2005.