Proceedings Paper:
Hughes, J and Atwell, E (1994) The automated evaluation of inferred word classifications. In: Cohn, AG, (ed.) ECAI-94 Proceedings of the 11th European Conference on Artificial Intelligence. ECAI-94: 11th European Conference on Artificial Intelligence, 08-12 Aug 1994, Amsterdam, The Netherlands. John Wiley & Sons, 535 - 539. ISBN: 0471950696.

The Automated Evaluation of Inferred Word Classifications

John Hughes and Eric Atwell[1]

Abstract.
Although automatically inferring classifications of words has been attempted by many researchers recently, no formal attempts to evaluate their results were made. Instead they relied on a *looks good to me* intuitive self-evaluation.
We outline a method by which automated word classification techniques can be fairly compared. The process by which words are automatically grouped into classes involves a number of decision points. The experiments selected a set of options for many of the decision points and rated each combination of the factors so that the most successful approach can be found. We directly compare some of the adopted approaches of other researchers with the set of factors that were found to produce the most linguistically plausible classification in our experiments. The evaluation method is also shown to be a valuable aid to highlighting approaches that are inefficient.

1 Hierarchical Clustering to Cluster Words

Hierarchical clustering is a way to produce a taxonomic classification of items such that, for a given cut-off point, the cut-off groups contain homogenous objects whilst the groups are as heterogeneous amongst themselves as possible. The items must have initially been compared with each other in such a way as there is a standard measure of similarity between each pair. The process begins by finding the closest two items and replacing them by a measurement which represents the *union* of the two in some meaningful way. Then, the second closest pair of items are searched for. This second group may consist of the first group merged with another item or it may consist of two new items. The items are collapsed in this way, iteratively, until all items become merged into the same group. As groups merge a record is kept of their similarity and a dendrogram forms. The method is described in further detail in [3] and some applications for this work are described in [4]. The choice of algorithm to calculate the distance of the two newly clustered items to all the other items as well as the distance metric to initially compare vectors can have a profound influence on the shape of the clustering. Each combination of metric and clustering method was tried in the experiments to see which derived the strongest syntactic classification of words in comparison to an intuitive linguistic classification (by the evaluation mechanisms described below). Three metrics were considered here: Manhattan, Euclidean and Spearman Rank Correlation Coefficient. The latter follows the modified definition given by [1] so that our results can be directly compared. Likewise, the choice of clustering method can greatly alter the resultant dendrogram after clustering; eight methods were included in the experiments described here, [7].

[1] Centre for Computer Analysis of Language and Speech, School of Computer Studies, Leeds University, Leeds LS2 9JT, UK

2 Automatic Evaluation

The last essential part of the automatic word classification process is some means of rating the quality of the alternative clustering schema for their accuracy. Other word classification projects have failed to include this vital procedure.

2.1 *Looks Good to Me*

Evaluating a clustering is typically done by the programmer using a *looks good to me* approach. To an extent he/she can feel how good one clustering is over another because he/she has an intrinsic understanding of the processes that produced it. However, he/she also has a vested interest in making his/her scheme look good. A more worthy evaluation can be done by an "independent" expert - in this case a linguist. It is rare to find one that has no bias in some way but his/her judgement based on experience must rate his/her appraisal above that of the programmer who has a vested interest to be seen to have done good work.
These evaluations are all done with some preconceived intuitive classification in mind. The actual question of what makes a good classification is not a simple one to answer. There are many alternatives and deciding which is superior comes down to personal judgement. Two rival clusterings may produce one winner when judged by one expert linguist but the other according to a different linguist's intuition. The linguist's intuition does not involve quantitative, measurable criteria, only qualitative overall impressions. The *looks good to me* approach may be fine if the aim is merely to demonstrate that patterns in text can classify words. This in itself is a laudable aim but if the best possible classification is desired then some way of comparing clustering schema is needed.

2.2 The LOB Benchmark Clustering

If it is accepted that a classification should conform closely to a syntactic intuitive one then there is a way it can be evaluated automatically thus resolving the problems of subjectivity amongst programmers and expert linguists. A *benchmark* classification can be derived which requires no input from the programmer nor a linguist but can be created empirically using a tagged corpus.

ECAI 94. *11th European Conference on Artificial Intelligence* Edited by A. Cohn
Published in 1994 by John Wiley & Sons, Ltd.

A benchmark was derived from the tagged LOB corpus using a broad, reduced tag-set. The novelty of the technique is that it yields a quantitative comparison against an existing benchmark. The LOB corpus provides an adequate source of tagging information but in principle the algorithm could equally be applied using another tagged corpus as a base.

To form the benchmark, firstly, counts of the assigned tags for each word are made. Words can now be compared to see how closely they coincide. Some words are only assigned one type of tag for every occurrence in the LOB corpus. Two unambiguous such words, classified with the same tag, will be clustered very closely. Next, words which are almost always of one particular tag are clustered close to the unambiguous words. In a similar way the more ambiguous words are classified according to their most common tag at a distance proportional to the degree of ambiguity. Words which share the same kind of ambiguity are classified very closely together.

The evaluation tool works by cutting the dendrogram at a certain point to produce a number of clusters. The members of the clusters can then be examined to see how they are tagged in the reduced LOB tag-set. A score can be calculated by classifying each group as the most common type amongst its members and counting up how many members conform to this type. A word that is tagged a noun more often than anything else will be judged to have been classified correctly if assigned to a group in which nouns dominate. A word that is a noun fairly frequently (the evaluator uses a threshold of 10% occurrence) but is tagged more frequently by at least one other type of tag in the LOB corpus will be judged to be *partially correct* and given a lower score. The score for each word reflects how consistent it is with the other words of the cluster it is situated in. A rating of the consitency of a cluster can be calculated by converting the sum of each member-word's score to a percentage. A cluster assigned a score of 100% would imply that every member-word has the same most frequent tag in LOB. By extension, the scores of individual words (*not* clusters) can be summed and converted to a percentage to give an overall rating of the quality of the clustering. The cut-point chosen will have a bearing on this process. A fair point is one that produces as few clusters as possible. The benchmark consists of 19 'reduced' tags such as *noun*, *past tense verb* and *cardinal number*. The dendrogram is cut at the point that produces 25 clusters which is very close to the ideal of 19 but still allows a little leeway. Deciding where to cut the dendrogram is obviously fairly ad hoc and other researchers in this area have ignored the issue altogether and arbitrarily chosen a cut off point that suits their purpose (usually a relatively large number of clusters to make their results look better). However, some of the experiments described here avoid the cut-point issue altogether by cutting the dendrogram at many points throughout its length. Two rival clustering schemes can then be contrasted by plotting graphs of the evaluations throughout the range of cut-points (see figure 1). A more complete description of the evaluation method is given in [3].

2.3 Automatically Evaluating Any Given Clustering

An alternative evaluation scheme does not use the benchmark but instead looks at the tagged LOB corpus to find how every word in the clustering is tagged. The rules follow from the benchmark used in the LOB experiments. Each word is compared with the LOB corpus to examine how it is tagged most often. The scoring regime follows that for the benchmark clust-ering. The overall score is calculated only for words present in the LOB corpus. The words *e-mail* and *email* in the example list below do not occur in LOB so are not included in any evaluation. An example of one of the least consistent groups from experiments to cluster 2000 words looks like this:

```
13) NOUN    85.3261%
.HALF      *DOG        *BRAND     *FIGURE     *REPLY     *DANCE
 ROUND     *KID         MIX       *ANSWER     *DEAL      *TRADE
*CHIP      *STEP       .SET       .OFFER      *CONTACT    DIE
*BOY       .DAMN       *SIGN      *GAIN       *TOUCH     *SLEEP
*DOCTOR    *FLAME      *WASTE     *PURCHASE    RESUME    .LIE
*BABY      *DREAM       e-mail    *POST       *DRINK     .FALL
*CHILD     *REQUEST     email      POSTING    *SWING     *VOTE
*CAT       *SURPRISE   *MAIL      *COMMENT    *DRESS     *WORK
```

If a word was tagged most frequently in LOB the same way as the tag assigned to its cluster (such as the majority of words in the example) then it was marked with a "*". If, instead, the second, third or fourth most common tag for the word in LOB matched its cluster's assigned tag (such as the words *half*, *damn*, *set*, *offer* *gain*, *lie* or *fall* in the example) then that word is marked with a ".". Words that do not match up (such as the words *round* *mix* or *die* in the example) aren't annotated at all. The words that are not present in LOB (*e-mail* and *email*) are printed in lower case whilst the recognised words are converted to upper case. The unknown words aren't included in any of the evaluation counts. A score out of 100 is calculated for each cluster using the same scoring methods for calculating an overall score. The example group was declared a **NOUN** group by the evaluator with approximately 85% accuracy.

3 Results

This section briefly records some of the results of various clustering schemes applied to some of the patterns in English language. The first set of experiments were carried out on a sample set of the 200 most frequent words in the LOB corpus as they appear in the untagged LOB corpus. The evaluation tool allows the best combination to be highlighted so that it can be used for much larger clusterings. An experiment to cluster 2000 words using the clustering method demonstrated to work best is also described.

3.1 Finding the Best Clustering Method

Table 1, below, contrasts the results for three distributional patterns formed by the position of a word in a sentence and two types of bigram counts. The notation $n \pm 2$ is used to indicate that bigrams were calculated for the number of times that the words to be classified co-occurred in the positions next-but-one before and next-but-one after a set of comparison items ($n \pm 1$ implies the positions immediately before and after, etc). The comparison items were the 101 most frequent lexical items in the LOB corpus. Normalized vectors were derived from statistics sampling the three patterns. Each combination of three metrics and eight clustering techniques were used to cluster the vectors (except for some of the third set of

experiments where results of certain combinations had already proved themselves not worthy of further investigation). The resultant dendrograms were evaluated, for the cut-off point where there were 25 clusters, against the benchmark clustering.

Table 1. Evaluations for 56 Clustering Experiments

Clustering Method	Positional Distribution Metric			Bigrams: $n \pm 1$ Metric			Bigrams: $n \pm 1$, $n \pm 2$ Metric		
	M	**E**	**S**	**M**	**E**	**S**	**M**	**E**	**S**
SL	25	29	23	38	31	29	–	–	–
CL	42	42	41	69	60	75	75	–	76
GA	38	37	36	72	46	74	70	–	69
WGA	40	41	41	73	50	70	74	–	71
Med	27	28	27	29	31	26	–	–	–
Cen	23	27	28	26	32	26	–	–	–
CoG	27	37	32	42	45	67	–	–	–
WM	43	45	42	76	64	74	79	–	77

- Clustering Methods
  - **SL** Single Linkage
  - **CL** Complete Linkage
  - **GA** Group Average
  - **WGA** Weighted Group Average
  - **Med** Median
  - **Cen** Centroid
  - **CoG** Centre of Gravity
  - **WM** Ward's Method
- Metrics
  - **M** Manhattan
  - **E** Euclidean
  - **S** Spearman Rank
- Results
  - – Indicates this experiment was not performed

The evaluations reveal that the context implied by sentence position distribution provides a poor representation of the syntactic rôle of the 200 words. The highest scoring combination consisting of the Euclidean metric and Ward's clustering method was only judged to be about 45% correct. The second set of experiments were made for bigram counts of the 200 most frequent words in the LOB corpus[2] appearing immediately before of after a target set of the most frequent 101 lexical items[3] in the corpus. This scored a great deal better than for the sentence position distribution. The highest scoring combination, Manhattan metric and Ward's clustering method scored 76%. Several experiments score much higher than the corresponding scores in the sentence position distribution experiments. The poor relative performance of sentence position distribution as a context measure meant it was not investigated further. However, there was clearly scope to investigate bigrams further. A third set of experiments, this time on just the best performing clustering schemes from the earlier experiments were carried out for bigrams covering the closest two neighbours on either side. These results are detailed on the right of the cells in Table 1. Figure 3 gives the full dendrogram, with the automatically assigned word-type for each cluster, of the highest scoring clustering scheme: Manhattan metric and Ward's method. The top left part of figure 3 lists the complete clustering scheme. The top right part shows the upper levels of the clustering above the point where the 25 clusters are cut. The dendrogram has been scaled from top-to-bottom by a factor of about ten times in order to display it neatly but the left-to-right measurements remain unaltered thus preserving the important dissimilarity information implicit in the dendrogram. The 25 clusters have been cut from the marked positions and displayed in the lower parts of the figure. They have not been scaled at all so relative dissimilarity measures are obvious. The labels for the groups are those that are assigned automatically by the evaluation program.

One factor of the experimental procedure which may have lead to false bias was the cut-off point at which the dendrogram was cut to form n clusters. Any bias due to the high dendrogram cut-off point used in the evaluator can be sidestepped if a graph is plotted for evaluations over a range of values. Figure 1 compares the highest scoring combination from our experiments, Manhattan metric and Ward's clustering method, with the combination that Finch believed to work best in his experiments, the Spearman Rank Correlation Coefficient metric and the Group Average clustering method.

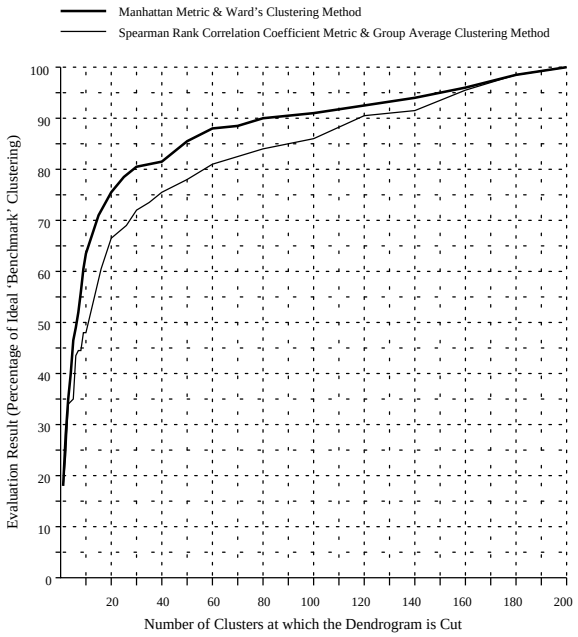


Figure 1. Comparing two Clustering Schemes

Clearly the combination of Manhattan metric and Ward's clustering method consistently outperforms the rival clustering scheme. Even when the dendrogram is cut at a very low level when the majority of the items are still singletons the combination of the Manhattan metric and Ward's clustering method scores higher and this advantage is retained the further one looks to the left along the graph rising to over a 10% advantage in some places. One feature of the Group Average clustering method that leads to its lesser performance is that often singletons remain unclustered until very late in the clustering process. The words *later*, *being* and *something* only get clustered when less than ten clusters remain using the Group Average method (with the Spearman Rank Correlation Coefficient metric). Ward's method is good at clustering these difficult "outliers" much earlier in the clustering process.

Another important factor affecting the clustering is the size of the comparison set (so far the same 101 lexical items have been used in the bigram experiments) The affect of varying

[2] The 200 words are those that appear in figure 3.
[3] This set includes punctuation.

this number was investigated next.
Ten experiments were carried out, each using ten more items in the comparison set than the previous one with the items being added in order of frequency. Thus, the third experiment used the 30 most frequent lexical items in the comparison set. The clustering scheme uses the Manhattan metric and Ward's clustering method. The results of these ten experiments are plotted in Figure 2.

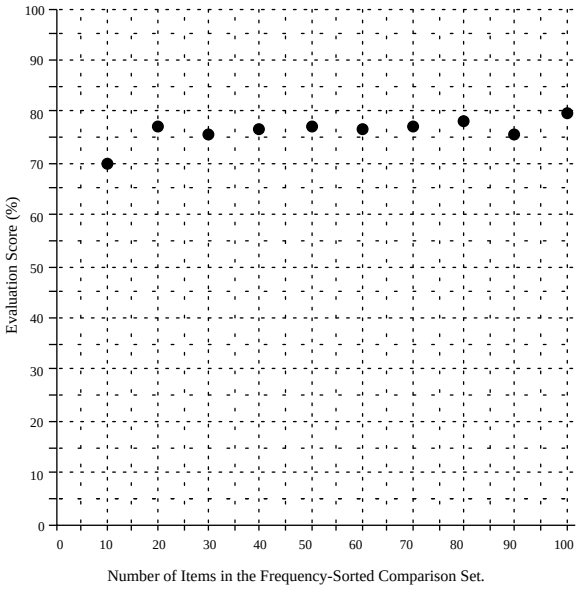


Figure 2. Evaluations for Comparison Sets of Varying Size

Just the ten most frequent lexical items lead to an evaluation of almost 70%. Adding in more and more items into the comparison set makes no significant difference to the quality of the clustering as measured by the evaluation tool. The reason the expressive power of these lexical items is so good is because they are mainly function words. [5] suggests that as these words are relatively unaffected by domain they act as markers for other words, hence indicating the categories of those words. So the experiment was classifying words by their relation to these function words which provide excellent contextual information. In Schütze's experiments to cluster 5000 words he used the context of bigram counts in the positions $n \pm 1$ and $n \pm 2$ as they co-occurred with the same 5000 words [6]. As the best contextual information seems to be provided by the function words - which make up the major part of the most frequent words in the corpus - it seems wasteful on resources to have such a large comparison set.

4 A Clustering of 2000 words

Now that the factors leading to a good clustering of words had been investigated we could select the best clustering scheme and use it to cluster a much larger set of words. The clustering scheme that used the distribution of bigram counts of the nearest four neighbours with a Manhattan metric and Ward's method was found to be most in line with intuitive expectations. We applied this scheme to a large corpus to cluster 2000 words.

For corpora of size 16 million and 35 million words the evaluations are very similar. When the dendrograms are cut at the point where there are 25 clusters (a very tightly constrained set for 2000 words) both scores are in the region of 80%. At the point where the dendrogram is cut to make 100 clusters the scores are in the region of 88% and for 400 clusters (corresponding the point where Finch cuts his dendrograms) the scores are around 94%. This implies that the corpus of 16 million words (a *third* the size of Finch's corpus) is representative of the bigram distribution and there is little to gain from using larger corpora.
The large-scale clustering was shown to not only group items of similar syntax but also to partially cluster items on their semantic or morphological similarity. When the dendrogram was cut to make 100 clusters the groups, listed below, resulted:

- Days Hours Minutes Weeks Months Years
- Feet Hands Fingers Eyes Legs Clothes Hair Arms Teeth Mind Opinion Chest Mouth Ass Breath Tongue Foot Arm Shoulder Face Head Heart Memory Name Voice
- Brother Sister Father Mother Daughter Son Mom Husband Wife
- Australia Canada America Europe Lebanon Vietnam California Cuba Boston Chicago
- Said Says Knows Feels Believes Thinks Assumed Believed Meant Claimed Stated Suggested Felt Knew Realized Figured Thought
- Keeping Having Buying Making Taking Giving Using Letting Adding Allowing Causing Leaving Bringing Putting Sending Finding
- David John Micheal Jack Bob Jim Brian Chris Dave Mike

Taken together, these results show that, although the clustering process is far from perfect, significant structure can be extracted from English without any prior knowledge of the domain and without supervision. Thus, empirical methods alone can uncover some language regularities. It remains to be shown how much of the structure of language can be uncovered with empiricist techniques.

REFERENCES

[1] S. Finch, *Finding Structure in Language*, PhD Thesis, Department of Cognitive Studies, Edinburgh University, 1993.

[2] J. Hughes and E. Atwell, Acquiring and Evaluating a Classification of Words. In S. Lucas, editor, *Grammatical Inference: Theory, Applications and Alternatives*, Colloqiuim at University of Essex, Colchester, 22^{nd} and 23^{rd} April 1993.

[3] J. Hughes, *Automatically Acquiring a Classification of Words*, PhD Thesis, School of Computer Studies, University of Leeds, 1994.

[4] J. Hughes and E. Atwell, A Methodical Approach to Word Class Formation Using Automatic Evaluation. To appear in L. Evett and T. Rose, editors, *Proceedings of AISB workshop on Computational Linguistics for Speech and Handwriting Recognition*, Leeds University, 12^{th} April 1994.

[5] D.M.W. Powers, On the Significance of Closed Classes and Boundary Conditions: Experiments in Lexical and Syntactic Learning. In W. Daelemans and D.M.W. Powers, editors, *Background and Experiments in Machine Learning of Natural Language*, Tilburg University, Institute for Language Technology and AI, 245-266, 1992.

[6] H. Schütze, *Part-of-Speech Induction from Scratch*, Technical Report, Centre for the Study of Language and Information, Stanford, 1993.

[7] J. Zupan, *Clustering of Large Data Sets*, John Wiley and Sons, Chichester, 1982.

Corpus: LOB

Clustered Items: Top 200 words

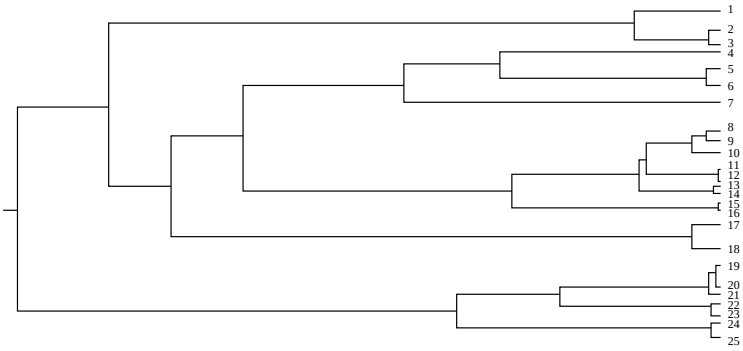
Comparison Set: Top 101 lexemes

Distribution: Bigram counts;
n-3, n+3 weighted 1
n-2, n+2 weighted 5
n-1, n+1 weighted 9

Metric: Manhattan

Clustering Method: Ward's

The dendrogram is cut to produce the 25 clusters shown below



1. VERB modal
'd
'll
can
could
may
must
should
will
might
would

2. VERB past
's
is
was
are
were
came
went
had
has
have

3. VERB present
did
does
do

4. PRONOUN
I
they
we
you
he
she
it
there

5. CONJUNCTION subordinating
after
before
since
until
as
that
because
if
when
but
though
where
while

6. WH-WORD
and
or
than
which
who

7. PREPOSITION
about
over
to
like
without
against
through
under
into
at
in
from
on
by
for
with
between
of
during

8. QUALIFIER
almost
no
being
quite
very
too
less
more

9. VERB present
be
been
come
go
going
right
something

10. ADVERB
also
still
always
not
never
n't

11. ADVERB
even
just
only
here
now
so
then
once
later
far
much
well

12. WH-WORD
how
what
why

13. VERB past
asked
told
said
thought

14. VERB past
found
made
left
put
given
used

15. VERB present
find
get
see
give
make
take

16. VERB present
know
think
say

17. ARTICLE
a
an
another
her
any
every
each
those
his
its
their
our
the
my
your

18. DETERMINER
Mr
these
this
all
both
such
many
some
most
one
three
two

19. ADVERB
London
fact
course
however

20. NOUN
case
time
way
day
year
man
government
world
hand
home
life
place
work
years
men
people

21. NOUN
number
part
use

22. ADVERB
again
away
back
out
down
up
off

23. PRONOUN
him
me
them
us
himself

24. ADJECTIVE
few
good
little
great
small
long

25. ADJECTIVE
first
last
same
new
young
old
other
own
per

Figure 3. The Highest Scoring Clustering in the LOB Experiments