Two Applications of Statistical Modelling to Natural Language Processing

William DuMouchel*, Carol Friedman, George Hripcsak, Stephen B Johnson, Paul D Clayton

Columbia University Center for Medical Informatics
161 Fort Washington Avenue, AP1310, New York, NY 10032
*Internet: dumouch@bayes.cpmc.columbia.edu

Abstract

Each week the Columbia-Presbyterian Medical Center collects several megabytes of English text transcribed from radiologists' dictation and notes of their interpretations of medical diagnostic x-rays. It is desired to automate the extraction of diagnoses from these natural language reports. This paper reports on two aspects of this project requiring advanced statistical methods. First, the identification of pairs of words and phrases that tend to appear together (collocate) uses a hierarchical Bayesian model that adjusts to different word and word pair distributions in different bodies of text. Second, we present an analysis of data from experiments to compare the performance of the computer diagnostic program to that of a panel of physician and lay readers of randomly sampled texts. A measure of inter-subject distance with respect to the diagnoses is defined for which estimated variances and covariances are easily computed. This allows statistical conclusions about the similarities and dissimilarities among diagnoses by the various programs and experts.

Empirical Bayes Estimation of Word Collocations

Friedman et al. (1994) describe a natural language processing (NLP) text extraction system, called MEDEXTRA, that was developed with the goal of becoming an integral component of the basic information needs of health care providers at Columbia Presbyterian Medical Center, a large health care facility. The general function of MEDEXTRA is the extraction, structuring and encoding of clinical information in textual patient reports, and the subsequent mapping of the information into a structured patient database which is used by other automated processes within the Clinical Information System (CIS), such as the decision support system or a research database. The first application of MEDEXTRA is to radiological reports, which are typically dictated by a radiologist and typed into the CIS by a clerk as unedited paragraphs of text. These reports typically deviate from the standard format because they are entered by many different typists, and all types of unpredictable variations eventually occur. Friedman et al (1994) describe the many components of MEDEXTRA, but this section focuses on a single aspect of the program, the development of a lexicon for multi-word phrases. The phrasal lexicon is critical to MEDEXTRA because the sublanguage is full of specialized expressions that should be treated as atomic units in order to obtain accurate interpretations. For example, the phrase *cannot be excluded,* as in *infiltrate cannot be excluded*, should convey the concept of *low possibility*.

The phrasal lexicon is constructed by a computer search for words that seem to occur together frequently, or *collocate*, after which a physician reviews the candidate phrases and recommends whether or not to include them in the lexicon. This section presents and compares three algorithms for the initial computer screening of word pairs. During analysis of a corpus of text, suppose that n separate word forms have been identified, and thus a potential of n^2 unique word pairs are to be examined for evidence of significant collocation. Let the proportion of times word form i is the first member of a pair be denoted by $p_{i.}$, $i = 1, ..., n$, and also let the proportion of times word form j is the second member of a pair be denoted by $p_{.j}$, $j = 1, ..., n$, where $\Sigma p_{i.} = \Sigma p_{.j} = 1$. Suppose that a sample of N word pairs have been gathered, and let N_{ij} denote the number of times word pair (i, j) has been observed, where $\Sigma_{i,j} N_{ij} = N$. Typically, N is large compared to n, but small compared to n^2, so that most of the $N_{ij} = 0$. For example, in a collection of mammogram reports, we find (see Table 1) $n = 2{,}043$, $N = 130{,}737$, and only 10,748 of the 4 million-plus $N_{ij} > 0$. We desire a measure of how much more frequently than chance a given pair of words arises, with error bars or some measure of statistical significance. Define $E_{ij} = N\, p_{i.}\, p_{.j} = N_{i.} N_{.j}/N$ as the expected number of occurrences of word pair (i, j) if the two words occur independently.

Fifth International Workshop on Artificial Intelligence and Statistics, Fort Lauderdale, FL, January 1995

The first measure of collocation considered is called the likelihood ratio or mutual information (MI) statistic. Dunning (1993) advocates this measure for assessing word collocation and points out its advantages over the simple Pearson chi-squared statistic when many of the observed counts are small. For pair (i, j) this measure is computed by comparing the observed and expected counts in the 2 by 2 table formed by classifying every observed word pair according to whether or not word i was first and/or word j was second. The likelihood-ratio test statistic for independence in the four-fold table is:

$$MI_{ij} = 2[N_{ij} \log \frac{N_{ij}}{E_{ij}} + (N_{i.}-N_{ij}) \log \frac{N_{i.} - N_{ij}}{N_{i.} - E_{ij}} + (N_{.j}-N_{ij}) \log \frac{N_{.j} - N_{ij}}{N_{.j} - E_{ij}}$$

$$+ (N-N_{i.}-N_{.j}+N_{ij}) \log \frac{N - N_{i.} - N_{.j} + N_{ij}}{N - N_{i.} - N_{.j} + E_{ij}}] \qquad (1)$$

The second measure is based on the simple observed proportion of times that word j follows word i, given that word i has come first, or vice-versa, whichever is larger. It is the maximum of two conditional probabilities (CP) and is computed as

$$CP_{ij} = N_{ij}/\min(N_{i.}, N_{.j}) \qquad \text{if } N_{ij} > 2, \text{ otherwise set } CP_{ij} = 0 \qquad (2)$$

Neither MI_{ij} nor CP_{ij} works well as a measure of word collocation. Since it is a test statistic and not an estimate of a population quantity, the mutual information statistic is quite dependent on sample size, and tends to be very large when there is only a moderate degree of collocation between words i and j but both $N_{i.}$ and $N_{.j}$ are large. The conditional proportion CP_{ij} tends to be large whenever one of $N_{i.}$ or $N_{.j}$ is much larger than the other. The requirement that N_{ij} be at least three is intended to reduce the effect of small frequencies on the conditional probabilities. The CP_{ij} measure is the one used in the version of MEDEXTRA described in Friedman et al (1994). Next we describe a new measure of collocation that seems to work better than either *MI* or *CP*. It focuses on identifying pairs in which the ratio $\lambda = E(N_{ij})/E_{ij}$ is large. To do this, we assume that each pair count N_{ij} has an independent Poisson distribution with mean $\lambda_{ij} E_{ij}$, written as

$$N_{ij} \sim \text{Poisson}(\lambda_{ij} E_{ij}) = \text{Poisson}(\lambda_{ij} N p_{i.} p_{.j})$$

$$P(N_{ij}=k \mid \lambda_{ij}=\lambda) = (\lambda E_{ij})^k e^{-\lambda E_{ij}} / k!$$

The value $\lambda_{ij} = 10$, for example, means that pair (i, j) occurs 10 times as frequently as would be expected by chance collocation. The goal is to estimate the larger values of λ_{ij} and produce confidence intervals for them. The main problem is that there are so many of the λ_{ij} to estimate that further model assumptions are necessary. We take a Bayesian approach to avoid the problem of *multiple comparisons*, whereby taking the largest of thousands of observed values of N_{ij}/E_{ij} at face value ignores the tendency of extreme values to regress toward the mean in future data.

The set of n^2 different λ_{ij} are assumed themselves to vary according to a probability distribution over the interval $0 < \lambda < \infty$. The exact form of this distribution is not known; however we assume that it is approximately a gamma distribution with parameters α and β, where these hyperparameters are estimated from the data. Under this assumption, the mean of all the λs is α/β, and the variance of λ is α/β^2. Thus, whatever the exact distribution of λ, proper choice of α and β can approximate its first two moments. Using Bayes rule, if the prior distribution is $\lambda_{ij} \sim \text{Gamma}(\alpha, \beta)$ and if the likelihood of the data is $N_{ij}|\lambda_{ij} \sim \text{Poisson}(\lambda_{ij}E_{ij})$ then the posterior distribution of λ_{ij} is also Gamma with revised parameters

$$\lambda_{ij}|N_{ij}, E_{ij} \sim \text{Gamma}(\alpha+N_{ij}, \beta+E_{ij})$$

The posterior mean and variance of λ_{ij} are $(\alpha+N_{ij})/(\beta+E_{ij})$ and $(\alpha+N_{ij})/(\beta+E_{ij})^2$, respectively. Arbitrary percentiles of the posterior distribution of λ_{ij} can be computed easily, using percentiles of the chi-squared distribution, which is a scaling of gamma distributions. For example, the first percentile of the posterior distribution is interpreted as the value of λ which one is 99% sure that $\lambda_{ij} = E(N_{ij})/E_{ij}$ exceeds, and could be considered a conservative Empirical Bayesian (EB) estimate (a lower confidence bound) of that quantity. It is

$$EB_{ij} = \chi^2_{.01}(2\alpha+2N_{ij})/(2\beta+2E_{ij}) \quad (3)$$

where $\chi^2_{.01}(\text{df})$ is the first percentile of the chi-squared distribution with df degrees of freedom. The values of α and β can be estimated from the marginal distribution of the N_{ij}. Although each N_{ij} has a Poisson distribution conditional on the corresponding λ_{ij}, unconditionally every N_{ij} has a negative binomial distribution that depends only on α, β and E_{ij}. The log-likelihood function is

$$\log L(\alpha, \beta) = \Sigma_{i,j} \{[\Sigma_{k=1, N_{ij}} \log(\alpha+k-1)] - N_{ij} \log(1 + \beta/E_{ij}) - \alpha \log(1 + E_{ij}/\beta)\} \quad (4)$$

The computation and maximization of $\log L(\alpha, \beta)$ is a delicate task when n is very large. Note that the first two terms of the summand in braces above vanish for all $N_{ij} = 0$, but that the last term must be evaluated for all n^2 pairs. The third measure of collocation used here is the quantity (3), where α and β maximize (4).

The resulting estimates should be more statistically reliable than using cutoffs based just on a chi-squared or mutual information criterion, or a simple conditional probability, since the estimation of α and β allows the method to adapt to the particular corpus being analyzed. Another advantage is that the interpretation of λ_{ij} is more straightforward than that of a test statistic — for example, its meaning is not dependent on the sample size. The Bayesian estimates are often called *shrinkage estimates*, because all the values N_{ij}/E_{ij} are "shrunk" towards α/β, which, because of the maximum likelihood estimation, will fall in the middle of the distribution of N_{ij}/E_{ij}. Extremely high and low values of N_{ij}/E_{ij} are thus automatically moderated.

The three methods are compared on three subdomains of the radiology text: abdominal x-rays, chest x-rays, and mammograms. The first three rows of Table 1 show statistics describing these three samples. The values of α and β are all in the range .01 to .02, indicating very high dispersion of λ in the superpopulation model. The fourth row of Table 1 describes the results using an artificial modification of the mammogram data in which all of the $N_{ij} > 2$ that were also greater than $10E_{ij}$ were reduced to $\max(2, 10E_{ij})$.

Table 2 (top) shows the correlation coefficients of the three measures across the 19,185 unique wordpairs occuring in the chest x-ray text. The correlations are small (all < .5) partly because the vast majority of the word pairs are not especially frequent. To give more emphasis to frequent collocations, and to balance the scales of the three measures (1) - (3), the three measures were replaced by their ranks (highest = 1, lowest = 19,185), and the logs of these ranks were correlated. The log scale is intended to ensure that the correlations are mostly determined by the pairs scoring high on the three measures. Table 2 (bottom) shows that the log-rank correlations range from .68 to .86 with the highest correlation between the Bayesian and mutual information measures. The same pattern held true for the other two subdomains studied.

Table 3 displays examples of the word pairs that each method is best and poorest at detecting. Within each of the three domains, a cutoff ranking of the three measures was based on the number of pairs for which $EB_{ij} > 10$. That is, according to their posterior distribution, we are 99% sure that such pairs are at least 10 times more likely to occur together than if they occurred independently. The last column of Table 1 shows how many word pairs met this criterion in each body of text. Then a cutoff score for the other two collocation measures was set so that the same number of pairs are chosen by each method. Table 3 shows examples of discordant word pairs -- the three measures do not all fall above or below their respective cutoffs. For example, at the top of Table 3, the phrase "treatment planning" ranked high on the MI measure but low on the CP and EB measures, while "or artifacts" ranked high on CP but not on MI or EB, and "chronic infection" ranked high on EB but not on the other two measures. Conversely, the first row of the second block in Table 3 shows that the phrase "coarse which" was not selected by MI, but was by both CP and EB, while "are too" was not selected by EB, although both MI and CP did select it. The phrases in Table 3 are samples of only 10 from each of their respective categories—there were usually between 200 and 500 in each category. But even these small samples are enough to show the pattern: the Bayesian choices tend to be more "interesting" medically, and the Bayesian omissions tend to be less interesting, than those in the other two columns of Table 3. The CP exclusive choices are cluttered by pairs containing common prepositions and conjunctions, while the MI exclusive choices are also biased toward choosing common words having slight tendencies to occur together. The Bayesian choices are better at focusing on reliably higher ratios of $E(N_{ij})/E_{ij}$ (> 10 with 99% confidence) because the Bayesian setup allows such word pairs to be explicitly described.

Finally, the last row of Table 1 shows results with the artificial modification of the mammogram text that reduced the frequency of pairs with a high ratio of N_{ij}/E_{ij}. The values of α and β increased by about a factor of 5, and the number of pairs having $EB_{ij} > 10$ dropped by a factor of 7. Although not shown for reasons of space, the same patterns as observed in Table 3 occur for this modified mammogram text. Another advantage of the Bayesian method is that the same cutoff score ($EB_{ij} > 10$) is reasonable for different populations and sample sizes. In order to restrict the choice to just 253 pairs in the modified mammogram text, the cutoff for *CP* must be increased from .18 in the original mammogram text to .57 and that for *MI* must be increased from 35 to 132. Cutoff levels for such measures are difficult to choose and somewhat arbitrary.

Estimating Distances Between Raters

As computers attempt to perform tasks, like those involving natural language, that attempt to mimic human expert performance, our evaluation of the computer's performance resembles a Turing test: we set up an experiment in which both the computer and human experts solve the same problems, and then see if a statistical analysis of the results can pick out the computer from among the humans. The analysis of such data is similar to an interrater reliability analysis, but the focus in not on estimating the overall consistency among raters or judges performing a task, but rather on how the computer's results look in comparison to both the average and the dispersion of the humans' results. One way to attack this problem is to estimate not just an overall interrater reliability score, but to estimate a distance measure between every pair of raters. If we can also estimate standard errors and a covariance matrix for these distance measures, then we will have the statistical tools to answer a variety of questions comparing the performances of the computer and of the human experts.

Hripcsak et al (1994) reports on such an experiment using the MEDEXTRA system. There were $n = 200$ independent test items to be assessed or rated, and each is rated by the NLP system (denoted as rater 0) and by some subset of J human experts (denoted as raters $j = 1, 2, ..., J$). Let X_{ij} be the rating score assigned to item i by judge j. Each X_{ij} may itself be a vector of scores, if each item is being rated on more than one characteristic. The experiment may not call on every judge to rate every item, so many of the X_{ij} may be missing at the time of the analysis. We will use subscripted n to denote how many items each judge or combinations of judges have rated. For example n_j denotes the number of items scored by judge j. Usually $n_0 = n$, if the computer rates all items, but this is not necessary to the analysis. Similarly, n_{jk} denotes the number of items rated by both judge j and judge k, n_{jklm} denotes the number of items rated by all of the judges j, k, l and m. There might be duplicates in the subscripts, in which case, for example, n_{jkjm} is interpreted as n_{jkm}.

Denote d_{ijk} to be some measure of distance between the rating scores, X_{ij} and X_{ik}, that judges j and k assign to item i. By convention, we set $d_{ijk} = 0$ if either of judges j or k did not rate item i. Let

$$\bar{d}_{jk} = \Sigma_i \, d_{ijk} / n_{jk}$$

$$V(\bar{d}_{jk}) = \mathrm{Cov}(\bar{d}_{jk}, \bar{d}_{jk})$$

$$\mathrm{Cov}(\bar{d}_{jk}, \bar{d}_{lm}) = [\Sigma_i \, d_{ijk} d_{ilm} - n_{jklm} \, \bar{d}_{jk} \bar{d}_{lm}] / n_{jk} n_{lm}$$

For matrix calculations, we can define the column vector $\bar{\boldsymbol{d}} = (\bar{d}_{01}, ..., \bar{d}_{J-1,J})^t$ of $J(J+1)/2$ mean distance scores, and the $J(J+1)/2$ by $J(J+1)/2$ covariance matrix $\boldsymbol{C}$, whose elements are defined above. The estimated variance of any linear combination $\boldsymbol{b}^t\bar{\boldsymbol{d}}$, where $\boldsymbol{b} = (b_{01}, ..., b_{J-1,J})$ is a vector of coefficients, is $V(\boldsymbol{b}^t\bar{\boldsymbol{d}}) = \boldsymbol{b}^t\boldsymbol{C}\boldsymbol{b}$.

The previous theory enables us to compute estimates and standard errors (and, assuming approximate normality, confidence intervals) for other measures of interrater difference. Some examples:

$\delta_j = \Sigma_{k \neq j}\, \bar{d}_{jk} / J \,;\, j = 0, 1, ..., J$ [Average distance of judge j from all other judges]

$\Delta = 2\Sigma_{k<j}\, \bar{d}_{jk} / J(J+1)$ [Overall average interjudge distance]

$\theta_j = \delta_0 - \delta_j \,;\, j = 1, 2, ..., J$ [Comparison of computer's mean distance with that of human judge j]

$\Sigma_j\, \theta_j / J$ [Comparison of computer's mean distance to average of all others' mean distances]

In the experiment reported by Hripcsak et al (1994), 18 human subjects and 3 automated algorithms are compared for their agreement in detecting a "reasonably likely" diagnosis of six different conditions from 200 randomly chosen chest x-ray reports. The six conditions are: congestive heart failure (CHF), chronic obstructive pulmonary disease, acute bacterial pneumonia, neoplasm, pleural effusion without CHF, and pneumothorax. The referenced paper provides details of the diagnoses and other aspects of the experimental design. The 18 human subjects included six radiologists, six internists, and six lay persons with no special medical training. Each human subject read 100 of the 200 reports, with each pair of subjects scoring at least 40 reports in common. To read a single report and choose among the six conditions took an average of 70 seconds for human subjects and 2 seconds for the natural language processor (running on a 42 Mhz IBM RS/6000 workstation). The primary computer algorithm of interest will be denoted NLP, and consists the combination of the MEDEXTRA program that reads reports and feeds data to an automated decision support system, and the decision support system which in turn draws conclusions and makes clinical recommendations. Two less sophisticated computer algorithms, based merely on keyword searches, were also included in the study, as well as the null algorithm that merely declares all six medical conditions absent from all reports. Each report was read by six of the twelve medical experts, and, assuming that a condition is present if a majority of experts (four or more out of six) voted for it, the prevalence of conditions ranged from 3% (chronic obstructive pulmonary disease) to 14% (acute bacterial pneumonia).

The distance measure used is the average fraction of diagnoses raters disagree on, and the primary outcome measure is the average distance of each subject to the (other) experts. The average distance of experts from each other was 0.24 (95% CI 0.19 - 0.29) Pairs of experts differed on the interpretation of reports for *at least* one diagnosis about 20% of the time. The average distance of NLP from the experts was 0.26 (95% CI 0.21 - 0.32). No two human experts were significantly more distant from the other experts than the average. The average distance of an expert to another expert of the same specialty (radiology or internal medicine) was almost exactly the same as the average distance of experts across specialties (0.24 vs 0.25, respectively). On the other hand, all of the six lay persons were much more distant from the experts. Their average distance to the 12 experts ranged from 0.51 to 0.73, with standard errors of about 0.03. The distance of the other automated algorithms from the experts was also significantly greater than the experts were from each other.

A multidimensional scaling analysis (Dillon and Goldstein, 1984) helps visualize the interrater distances. Twenty-one raters (the poorest performing keyword search algorithm was excluded) are represented by 21 points in the plane that attempt to preserve the interrater distances $\bar{d}_{jk}$. The Figure displays the results. The 12 experts and the NLP cluster in the lower left of the Figure, while the lay persons are placed far to the right along with the null algorithm, and the complex keyword search stands alone at the upper left of the Figure. The Hripcsak et al report shows other graphical representations of these data, including a sensitivity-specificity plot of all raters. This shows that dimension 1 of the present multidimensional scaling analysis is primarily variation in sensitivity (the null algorithm has zero sensitivity, of course), and dimension 2 is primarily variation in specificity (the keyword search algorithm and one of the lay persons have many false positives).

In conclusion, these two examples show that statistical modelling can help develop a more focused approach to developing and evaluating natural language processing algorithms. The Bayesian word collocation analysis provides a more relevent measure of collocation and thus more discrimination among potential word pairs. In the evaluation experiment, the ability to get standard errors and confidence intervals for every interrater distance, and, more importantly, for all linear combinations of these distances, provides convincing evidence of the power of the NLP algorithm to work well in practice.

References

Dillon W, Goldstein M (1984) *Multivariate Analysis*, New York: Wiley, 587pp.

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

Friedman C, Hripcsak G, DuMouchel W, Johnson S, Clayton P (1994) Natural language processing in an operational clinical information system (*Submitted for publication*)

Hripcsak G, Friedman C, Alderson P, DuMouchel W, Johnson S, Clayton P (1994) Unlocking clinical data from narrative reports (*Submitted for publication*)

Domain	n	$N_{ij} > 0$	N	α ± st.err.	β ± st.err.	$Eb_{ij} > 10$
Abdomen	3,542	21,058	110,415	.0174 ± .0001	.0170 ± .0002	2205
Chest	3,274	19,185	128,924	.0179 ± .0001	.0168 ± .0003	2029
Mammogram	2,043	10,748	130,737	.0133 ± .0001	.0096 ± .0002	1766
Modified Mamm.	2,043	10,748	69,167	.0601 ± .0006	.0533 ± .0013	253

Table 1. Statistics for each of the bodies of text.

	MI	CP	EB
MI	1.000	0.287	0.269
CP	0.287	1.000	0.477
EB	0.269	0.477	1.000

	lrMI	lrCP	lrEB
lrMI	1.000	0.677	0.861
lrCP	0.677	1.000	0.757
lrEB	0.861	0.757	1.000

Table 2. Correlations among the three measures (top) and among their log ranks (bottom) on the chest x-ray text.

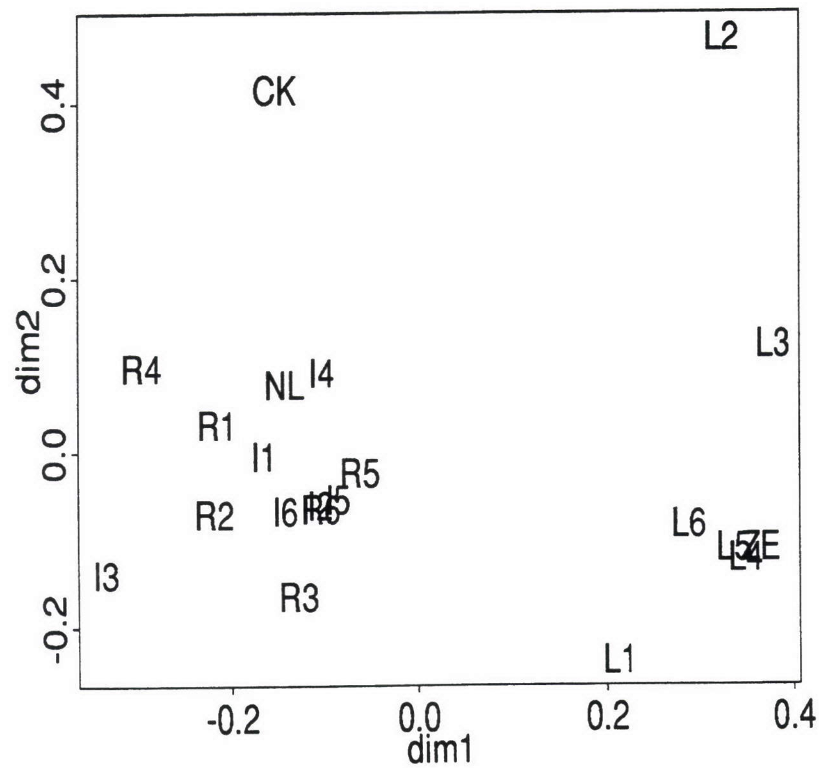

Figure (right). Location of each "subject" in a two-dimensional MDS fit to the between-rater distances. (Internists: I1...I6, Radiologists: R1...R6, Lay persons: L1...L6, Natural Lang. Proc.: NL, Complex Keyword: CK, All 0: ZE.)

Mutual Information	**Max Conditional Prob.**	**Bayesian Posterior Dist.**
Abdomen X-ray Domain -- Exclusively Chosen:		
treatment planning	or artifacts	chronic infection
floor relaxation	impacted in	represent chronic
portable abdominal	is perhaps	multiple mobile
easily palpable	left paravertebral	irregular calcification
pattern appears	of ureteropelvic	some minor
films are	in lumbosacral	loculated effusion
emergency room	or benign	suggest gallstone
normal abdominal	site is	simple appearing
to right	to abscence	markedly limited
and supine	within ovary	patent splenic
Abdomen X-ray Domain -- Exclusively Omitted:		
coarse which	lateral segment	are too
remain clear	nodular densities	of prostate
poor visualization	at 12	diameter of
abnormal endocrinologic	it measures	dimension and
subdiaphragmatic region	radiographs were	from <*NUMBER*>
large hemorrhagic	previous films	pa and
ngt tip	nondilated air	is not
obstructing lesion	pelvic mass	is unremarkable
which correlates	masses can	abdomen and
prior dictation	prostate carcinoma	is normal
Chest X-ray Domain -- Exclusively Chosen:		
night sweats	the pedicles	suggested when
the pneumothorax	is midline	exclude bibasilar
lungs appear	is much	only minimal
obtained in	however the	correlation recommended
and this	be partially	nd through
mild to	origin of	mm metallic
studies are	and ectatic	bibasilar haziness
and are	layering of	important bony
which probably	left humerus	similar examination
house staff	of copd	aortic tortuosity
Chest X-ray Domain -- Exclusively Omitted:		
from september	lower neck	dome of
description :	subclavian vein	the same
somewhat unusual	enlarged but	is centrally
multiple lucencies	defined density	the aortic
operatively demonstrates	cardiac pathology	is identified
continues to	exam :	the trachea
out aspiration	probably due	of both
do not	pericardial pathology	the thoracic
substantial change	rounded opacity	the endotracheal
+)	pericardial clips	to the
Mammography Domain -- Exclusively Chosen:		
the amount	the eight	dr manson
breasts the	breast ranging	also demonstrated
breast of	showed no	discussed extensively
the mammography	a unilateral	irregular suggestive
or of	minimal to	patient complained
the mammographic	is once	clearly visualized
the microcalcifications	involution of	number when
or parenchyma	a pacemaker	s injury
a symmetric	had a	postradiation changes
scattered microcalcifications	of glandular	lower central
Mammography Domain -- Exclusively Omitted:		
while this	metropolitan hospital	of residual
office was	was indeterminate	which time
be done	changed when	the periareolar
calcifications many	reduction mammoplasty	the retroareolar
was introduced	eosinophilic pneumonia	is suggested
fluctuating cyst	years ago	are two
patient return	suggested particularly	a rounded
left axial	post surgical	the site
completely unchanged	appears mammographically	is an
was notified	cm from	is moderately

Table 3. Samples of phrases exclusively chosen and exclusively omitted by each of the three methods.