

Concept Frequency Distribution in Biomedical Text Summarization

Lawrence H. Reeve¹, Hyoil Han¹, Saya V. Nagori², Jonathan C. Yang², Tamara A. Schwimmer², Ari D. Brooks²

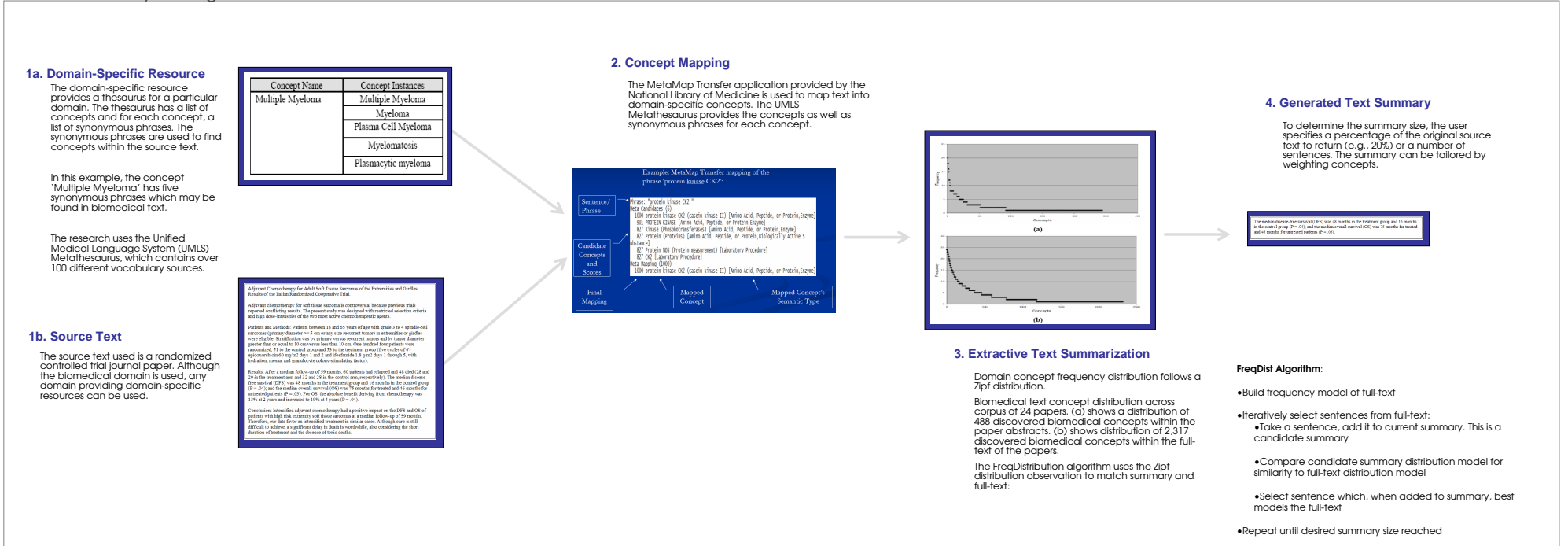
¹ Drexel University, College of Information Science and Technology
² Drexel University, College of Medicine

Abstract

Text summarization is a data reduction process. The use of text summarization enables users to reduce the amount of text that must be read while still assimilating the core information. The data reduction offered by text summarization is particularly useful in the biomedical domain, where physicians must continuously find clinical trial study information to incorporate into their patient treatment efforts. Such efforts are often hampered by the high-volume of publications.

Our contribution is two-fold: 1) to propose the frequency of domain concepts as a method to identify important sentences within a full-text; and 2) propose a new model and algorithm for identifying important sentences based on term or concept frequency distribution. It is shown that the use of concepts performs closely with the use of terms for sentence selection. Our proposed frequency distribution model algorithm outperforms a state-of-the-art approach.

Figure 1: Text Summarization from Concepts Using Frequency Distribution



Motivation, Hypothesis and Method

Motivation: Generate short summaries of biomedical texts (randomized controlled trials in oncology) to allow physicians and researchers to assimilate more information in less time.

Approach: Identify and extract sentences from the source text to form a summary. The problem is how to identify important sentences within the full source text to extract while simultaneously reducing information redundancy.

Hypothesis: The distribution of concepts within a summary should approximate the same distribution of concepts in the full source.

Why not use the abstract?

- No ideal summary exists
- Abstract is one summary from one viewpoint
 - Doesn't consider user's information need
- Abstract may be missing content from full-text
- Generation of customized content
 - Question-answering systems
- Evaluate sentence selection methods

Method: Build a summary incrementally by adding a sentence to the summary and then comparing the summary concept distribution to the source text concept distribution. To compare similarity between summary and source concept distribution, the distributions for each are modeled as vectors, and then compared using the methods below:

$$\text{cosine} = \frac{\sum_{i=1}^n \min(x_i, y_i)}{\sum_{i=1}^n \max(x_i, y_i)}$$

$$\text{Dice} = \frac{2 \times \sum_{i=1}^n \min(x_i, y_i)}{\sum_{i=1}^n (x_i + y_i)}$$

$$\text{Jaccard} = \frac{\sum_{i=1}^n \min(x_i, y_i)}{\sum_{i=1}^n (\max(x_i, y_i) - \min(x_i, y_i))}$$

$$\text{Dice} = \frac{2 \times \sum_{i=1}^n \min(x_i, y_i)}{\sum_{i=1}^n (x_i + y_i)}$$

$$\text{Dice} = \frac{2 \times \sum_{i=1}^n \min(x_i, y_i)}{\sum_{i=1}^n (x_i + y_i)}$$

(b) Vector similarity
 Sparsity weighting to reduce a candidate summary's semantic distance to the target using the concept distribution (a) cosine similarity (b) Dice coefficient (c) Jaccard distance (d) skip bigram (e) skip bigram (f) skip bigram (g) skip bigram (h) skip bigram (i) skip bigram (j) skip bigram (k) skip bigram (l) skip bigram (m) skip bigram (n) skip bigram (o) skip bigram (p) skip bigram (q) skip bigram (r) skip bigram (s) skip bigram (t) skip bigram (u) skip bigram (v) skip bigram (w) skip bigram (x) skip bigram (y) skip bigram (z) skip bigram (aa) skip bigram (ab) skip bigram (ac) skip bigram (ad) skip bigram (ae) skip bigram (af) skip bigram (ag) skip bigram (ah) skip bigram (ai) skip bigram (aj) skip bigram (ak) skip bigram (al) skip bigram (am) skip bigram (an) skip bigram (ao) skip bigram (ap) skip bigram (aq) skip bigram (ar) skip bigram (as) skip bigram (at) skip bigram (au) skip bigram (av) skip bigram (aw) skip bigram (ax) skip bigram (ay) skip bigram (az) skip bigram (ba) skip bigram (bb) skip bigram (bc) skip bigram (bd) skip bigram (be) skip bigram (bf) skip bigram (bg) skip bigram (bh) skip bigram (bi) skip bigram (bj) skip bigram (bk) skip bigram (bl) skip bigram (bm) skip bigram (bn) skip bigram (bo) skip bigram (bp) skip bigram (bq) skip bigram (br) skip bigram (bs) skip bigram (bt) skip bigram (bu) skip bigram (bv) skip bigram (bw) skip bigram (bx) skip bigram (by) skip bigram (bz) skip bigram (ca) skip bigram (cb) skip bigram (cc) skip bigram (cd) skip bigram (ce) skip bigram (cf) skip bigram (cg) skip bigram (ch) skip bigram (ci) skip bigram (cj) skip bigram (ck) skip bigram (cl) skip bigram (cm) skip bigram (cn) skip bigram (co) skip bigram (cp) skip bigram (cq) skip bigram (cr) skip bigram (cs) skip bigram (ct) skip bigram (cu) skip bigram (cv) skip bigram (cw) skip bigram (cx) skip bigram (cy) skip bigram (cz) skip bigram (da) skip bigram (db) skip bigram (dc) skip bigram (dd) skip bigram (de) skip bigram (df) skip bigram (dg) skip bigram (dh) skip bigram (di) skip bigram (dj) skip bigram (dk) skip bigram (dl) skip bigram (dm) skip bigram (dn) skip bigram (do) skip bigram (dp) skip bigram (dq) skip bigram (dr) skip bigram (ds) skip bigram (dt) skip bigram (du) skip bigram (dv) skip bigram (dw) skip bigram (dx) skip bigram (dy) skip bigram (dz) skip bigram (ea) skip bigram (eb) skip bigram (ec) skip bigram (ed) skip bigram (ee) skip bigram (ef) skip bigram (eg) skip bigram (eh) skip bigram (ei) skip bigram (ej) skip bigram (ek) skip bigram (el) skip bigram (em) skip bigram (en) skip bigram (eo) skip bigram (ep) skip bigram (eq) skip bigram (er) skip bigram (es) skip bigram (et) skip bigram (eu) skip bigram (ev) skip bigram (ew) skip bigram (ex) skip bigram (ey) skip bigram (ez) skip bigram (fa) skip bigram (fb) skip bigram (fc) skip bigram (fd) skip bigram (fe) skip bigram (ff) skip bigram (fg) skip bigram (fh) skip bigram (fi) skip bigram (fj) skip bigram (fk) skip bigram (fl) skip bigram (fm) skip bigram (fn) skip bigram (fo) skip bigram (fp) skip bigram (fq) skip bigram (fr) skip bigram (fs) skip bigram (ft) skip bigram (fu) skip bigram (fv) skip bigram (fw) skip bigram (fx) skip bigram (fy) skip bigram (fz) skip bigram (ga) skip bigram (gb) skip bigram (gc) skip bigram (gd) skip bigram (ge) skip bigram (gf) skip bigram (gg) skip bigram (gh) skip bigram (gi) skip bigram (gj) skip bigram (gk) skip bigram (gl) skip bigram (gm) skip bigram (gn) skip bigram (go) skip bigram (gp) skip bigram (gq) skip bigram (gr) skip bigram (gs) skip bigram (gt) skip bigram (gu) skip bigram (gv) skip bigram (gw) skip bigram (gx) skip bigram (gy) skip bigram (gz) skip bigram (ha) skip bigram (hb) skip bigram (hc) skip bigram (hd) skip bigram (he) skip bigram (hf) skip bigram (hg) skip bigram (hh) skip bigram (hi) skip bigram (hj) skip bigram (hk) skip bigram (hl) skip bigram (hm) skip bigram (hn) skip bigram (ho) skip bigram (hp) skip bigram (hq) skip bigram (hr) skip bigram (hs) skip bigram (ht) skip bigram (hu) skip bigram (hv) skip bigram (hw) skip bigram (hx) skip bigram (hy) skip bigram (hz) skip bigram (ia) skip bigram (ib) skip bigram (ic) skip bigram (id) skip bigram (ie) skip bigram (if) skip bigram (ig) skip bigram (ih) skip bigram (ii) skip bigram (ij) skip bigram (ik) skip bigram (il) skip bigram (im) skip bigram (in) skip bigram (io) skip bigram (ip) skip bigram (iq) skip bigram (ir) skip bigram (is) skip bigram (it) skip bigram (iu) skip bigram (iv) skip bigram (iw) skip bigram (ix) skip bigram (iy) skip bigram (iz) skip bigram (ja) skip bigram (jb) skip bigram (jc) skip bigram (jd) skip bigram (je) skip bigram (jf) skip bigram (jg) skip bigram (jh) skip bigram (ji) skip bigram (jj) skip bigram (jk) skip bigram (jl) skip bigram (jm) skip bigram (jn) skip bigram (jo) skip bigram (jp) skip bigram (jq) skip bigram (jr) skip bigram (js) skip bigram (jt) skip bigram (ju) skip bigram (jv) skip bigram (jw) skip bigram (jx) skip bigram (jy) skip bigram (jz) skip bigram (ka) skip bigram (kb) skip bigram (kc) skip bigram (kd) skip bigram (ke) skip bigram (kf) skip bigram (kg) skip bigram (kh) skip bigram (ki) skip bigram (kj) skip bigram (kl) skip bigram (km) skip bigram (kn) skip bigram (ko) skip bigram (kp) skip bigram (kq) skip bigram (kr) skip bigram (ks) skip bigram (kt) skip bigram (ku) skip bigram (kv) skip bigram (kw) skip bigram (kx) skip bigram (ky) skip bigram (kz) skip bigram (la) skip bigram (lb) skip bigram (lc) skip bigram (ld) skip bigram (le) skip bigram (lf) skip bigram (lg) skip bigram (lh) skip bigram (li) skip bigram (lj) skip bigram (lk) skip bigram (ll) skip bigram (lm) skip bigram (ln) skip bigram (lo) skip bigram (lp) skip bigram (lq) skip bigram (lr) skip bigram (ls) skip bigram (lt) skip bigram (lu) skip bigram (lv) skip bigram (lw) skip bigram (lx) skip bigram (ly) skip bigram (lz) skip bigram (ma) skip bigram (mb) skip bigram (mc) skip bigram (md) skip bigram (me) skip bigram (mf) skip bigram (mg) skip bigram (mh) skip bigram (mi) skip bigram (mj) skip bigram (mk) skip bigram (ml) skip bigram (mm) skip bigram (mn) skip bigram (mo) skip bigram (mp) skip bigram (mq) skip bigram (mr) skip bigram (ms) skip bigram (mt) skip bigram (mu) skip bigram (mv) skip bigram (mw) skip bigram (mx) skip bigram (my) skip bigram (mz) skip bigram (na) skip bigram (nb) skip bigram (nc) skip bigram (nd) skip bigram (ne) skip bigram (nf) skip bigram (ng) skip bigram (nh) skip bigram (ni) skip bigram (nj) skip bigram (nk) skip bigram (nl) skip bigram (nm) skip bigram (nn) skip bigram (no) skip bigram (np) skip bigram (nq) skip bigram (nr) skip bigram (ns) skip bigram (nt) skip bigram (nu) skip bigram (nv) skip bigram (nw) skip bigram (nx) skip bigram (ny) skip bigram (nz) skip bigram (oa) skip bigram (ob) skip bigram (oc) skip bigram (od) skip bigram (oe) skip bigram (of) skip bigram (og) skip bigram (oh) skip bigram (oi) skip bigram (oj) skip bigram (ok) skip bigram (ol) skip bigram (om) skip bigram (on) skip bigram (oo) skip bigram (op) skip bigram (oq) skip bigram (or) skip bigram (os) skip bigram (ot) skip bigram (ou) skip bigram (ov) skip bigram (ow) skip bigram (ox) skip bigram (oy) skip bigram (oz) skip bigram (pa) skip bigram (pb) skip bigram (pc) skip bigram (pd) skip bigram (pe) skip bigram (pf) skip bigram (pg) skip bigram (ph) skip bigram (pi) skip bigram (pj) skip bigram (pk) skip bigram (pl) skip bigram (pm) skip bigram (pn) skip bigram (po) skip bigram (pp) skip bigram (pq) skip bigram (pr) skip bigram (ps) skip bigram (pt) skip bigram (pu) skip bigram (pv) skip bigram (pw) skip bigram (px) skip bigram (py) skip bigram (pz) skip bigram (qa) skip bigram (qb) skip bigram (qc) skip bigram (qd) skip bigram (qe) skip bigram (qf) skip bigram (qg) skip bigram (qh) skip bigram (qi) skip bigram (qj) skip bigram (qk) skip bigram (ql) skip bigram (qm) skip bigram (qn) skip bigram (qo) skip bigram (qp) skip bigram (qq) skip bigram (qr) skip bigram (qs) skip bigram (qt) skip bigram (qu) skip bigram (qv) skip bigram (qw) skip bigram (qx) skip bigram (qy) skip bigram (qz) skip bigram (ra) skip bigram (rb) skip bigram (rc) skip bigram (rd) skip bigram (re) skip bigram (rf) skip bigram (rg) skip bigram (rh) skip bigram (ri) skip bigram (rj) skip bigram (rk) skip bigram (rl) skip bigram (rm) skip bigram (rn) skip bigram (ro) skip bigram (rp) skip bigram (rq) skip bigram (rr) skip bigram (rs) skip bigram (rt) skip bigram (ru) skip bigram (rv) skip bigram (rw) skip bigram (rx) skip bigram (ry) skip bigram (rz) skip bigram (sa) skip bigram (sb) skip bigram (sc) skip bigram (sd) skip bigram (se) skip bigram (sf) skip bigram (sg) skip bigram (sh) skip bigram (si) skip bigram (sj) skip bigram (sk) skip bigram (sl) skip bigram (sm) skip bigram (sn) skip bigram (so) skip bigram (sp) skip bigram (sq) skip bigram (sr) skip bigram (ss) skip bigram (st) skip bigram (su) skip bigram (sv) skip bigram (sw) skip bigram (sx) skip bigram (sy) skip bigram (sz) skip bigram (ta) skip bigram (tb) skip bigram (tc) skip bigram (td) skip bigram (te) skip bigram (tf) skip bigram (tg) skip bigram (th) skip bigram (ti) skip bigram (tj) skip bigram (tk) skip bigram (tl) skip bigram (tm) skip bigram (tn) skip bigram (to) skip bigram (tp) skip bigram (tq) skip bigram (tr) skip bigram (ts) skip bigram (tt) skip bigram (tu) skip bigram (tv) skip bigram (tw) skip bigram (tx) skip bigram (ty) skip bigram (tz) skip bigram (ua) skip bigram (ub) skip bigram (uc) skip bigram (ud) skip bigram (ue) skip bigram (uf) skip bigram (ug) skip bigram (uh) skip bigram (ui) skip bigram (uj) skip bigram (uk) skip bigram (ul) skip bigram (um) skip bigram (un) skip bigram (uo) skip bigram (up) skip bigram (uq) skip bigram (ur) skip bigram (us) skip bigram (ut) skip bigram (uu) skip bigram (uv) skip bigram (uw) skip bigram (ux) skip bigram (uy) skip bigram (uz) skip bigram (va) skip bigram (vb) skip bigram (vc) skip bigram (vd) skip bigram (ve) skip bigram (vf) skip bigram (vg) skip bigram (vh) skip bigram (vi) skip bigram (vj) skip bigram (vk) skip bigram (vl) skip bigram (vm) skip bigram (vn) skip bigram (vo) skip bigram (vp) skip bigram (vq) skip bigram (vr) skip bigram (vs) skip bigram (vt) skip bigram (vu) skip bigram (vv) skip bigram (vw) skip bigram (vx) skip bigram (vy) skip bigram (vz) skip bigram (wa) skip bigram (wb) skip bigram (wc) skip bigram (wd) skip bigram (we) skip bigram (wf) skip bigram (wg) skip bigram (wh) skip bigram (wi) skip bigram (wj) skip bigram (wk) skip bigram (wl) skip bigram (wm) skip bigram (wn) skip bigram (wo) skip bigram (wp) skip bigram (wq) skip bigram (wr) skip bigram (ws) skip bigram (wt) skip bigram (wu) skip bigram (wv) skip bigram (ww) skip bigram (wx) skip bigram (wy) skip bigram (wz) skip bigram (xa) skip bigram (xb) skip bigram (xc) skip bigram (xd) skip bigram (xe) skip bigram (xf) skip bigram (xg) skip bigram (xh) skip bigram (xi) skip bigram (xj) skip bigram (xk) skip bigram (xl) skip bigram (xm) skip bigram (xn) skip bigram (xo) skip bigram (xp) skip bigram (xq) skip bigram (xr) skip bigram (xs) skip bigram (xt) skip bigram (xu) skip bigram (xv) skip bigram (xw) skip bigram (xx) skip bigram (xy) skip bigram (xz) skip bigram (ya) skip bigram (yb) skip bigram (yc) skip bigram (yd) skip bigram (ye) skip bigram (yf) skip bigram (yg) skip bigram (yh) skip bigram (yi) skip bigram (yj) skip bigram (yk) skip bigram (yl) skip bigram (ym) skip bigram (yn) skip bigram (yo) skip bigram (yp) skip bigram (yq) skip bigram (yr) skip bigram (ys) skip bigram (yt) skip bigram (yu) skip bigram (yv) skip bigram (yw) skip bigram (yx) skip bigram (yy) skip bigram (yz) skip bigram (za) skip bigram (zb) skip bigram (zc) skip bigram (zd) skip bigram (ze) skip bigram (zf) skip bigram (zg) skip bigram (zh) skip bigram (zi) skip bigram (zj) skip bigram (zk) skip bigram (zl) skip bigram (zm) skip bigram (zn) skip bigram (zo) skip bigram (zp) skip bigram (zq) skip bigram (zr) skip bigram (zs) skip bigram (zt) skip bigram (zu) skip bigram (zv) skip bigram (zw) skip bigram (zx) skip bigram (zy) skip bigram (zz)

Figure 2: Concept Similarity Functions evaluated

Evaluation

Basic Idea

- Generate ideal ("model") summaries from domain experts
- Generate system summaries and compare to model summaries

Resources

- A corpus of 24 randomly selected full biomedical texts was used
- Three domain experts generated extractive model summaries for each paper (20% compression)
- Six system summarizers generated extractive summaries (20% compression), used for comparison

ROUGE n-gram overlap tool

- ROUGE: Recall Oriented Understudy for Gisting Evaluation
- Compares system summaries to model summaries
- Results based on n-gram overlap
- ROUGE-2: bigram co-occurrence
- ROUGE-SU4: skip bigram with no more than 4 intervening words
- Same metrics as used in 2005 Document Understanding Conference

Results & Discussion

ROUGE-2 Scores		ROUGE-SU4 Scores	
FrqDist-24	0.171	FrqDist-24	0.101
FrqDist-24-20	0.164	FrqDist-24-20	0.100
FrqDist-24-10	0.161	FrqDist-24-10	0.100
FrqDist-24-5	0.159	FrqDist-24-5	0.100
FrqDist-24-2	0.157	FrqDist-24-2	0.100
FrqDist-24-1	0.155	FrqDist-24-1	0.100
FrqDist-24-0	0.153	FrqDist-24-0	0.100
FrqDist-24-0.5	0.151	FrqDist-24-0.5	0.100
FrqDist-24-0.2	0.149	FrqDist-24-0.2	0.100
FrqDist-24-0.1	0.147	FrqDist-24-0.1	0.100
FrqDist-24-0.05	0.145	FrqDist-24-0.05	0.100
FrqDist-24-0.02	0.143	FrqDist-24-0.02	0.100
FrqDist-24-0.01	0.141	FrqDist-24-0.01	0.100
FrqDist-24-0.005	0.139	FrqDist-24-0.005	0.100
FrqDist-24-0.002	0.137	FrqDist-24-0.002	0.100
FrqDist-24-0.001	0.135	FrqDist-24-0.001	0.100
FrqDist-24-0.0005	0.133	FrqDist-24-0.0005	0.100
FrqDist-24-0.0002	0.131	FrqDist-24-0.0002	0.100
FrqDist-24-0.0001	0.129	FrqDist-24-0.0001	0.100
FrqDist-24-0.00005	0.127	FrqDist-24-0.00005	0.100
FrqDist-24-0.00002	0.125	FrqDist-24-0.00002	0.100
FrqDist-24-0.00001	0.123	FrqDist-24-0.00001	0.100

Figure 3: ROUGE Scores

Observations:

- Random summarizer performs relatively well!
- Ignoring FreqDist, Baseline Lead is worst performing
 - Unlike news genre, where Lead is competitive
- Use of Dice for FreqDist works best
- Use of terms and concepts perform closely
 - but concepts can allow for easier tailoring of a summary