A Note on CLARIT 'Similarity' Performance

David A. Evans, Steve K. Handerson, Robert G. Lefferts
Ira A. Monarch, William R. Hersh

December 1991

Report No. CMU-LCL-91-3

Laboratory for Computational Linguistics

139 Baker Hall
Department of Philosophy
Carnegie Mellon University
Pittsburgh, PA 15213

A Note on CLARIT 'Similarity' Performance

Laboratory for Computational Linguistics
Carnegie Mellon University
David A. Evans, Steve K. Handerson, Robert G. Lefferts
Ira A. Monarch, William R. Hersh

December 1991

Contents

1 Introduction

A recent experiment (Hersh et al. 1991) contrasted CLARIT indexing with word-based and non-NLP-phrase-based approaches. The task involved identifying previously-determined *similar* vs. *dissimilar* pairs of documents in a full-text database. The results showed that CLARIT-based indexing clearly outperformed the other two approaches in document discrimination. In the work reported here, the procedure was repeated with word-based processing and CLARIT. In this case, though not in the original experiment, CLARIT-nominated terms were used along with the words they contained, thus affording 'partial' matches of terms.

2 Method

Forty-two, full-text articles from the *New England Journal of Medicine* dealing with *AIDS* were chosen by a physician colleague (Hersh). A resident was asked to read the articles and identify 'similar' and 'dissimilar' articles. This resulted in six pairs of articles that the resident judged to be highly 'similar' and two pairs of articles that the resident judged to be highly 'dissimilar'. (The target articles are given in Appendix A.)

All forty-two articles were indexed by a word-based indexing program, SWORD, and by CLARIT. SWORD is modeled on Salton's SMART system; indexing is based on stemmed words (not including 'stop-listed' items) which are weighted for relevance. In SWORD indexing, 'relevance' is given by a word's *IDF*TF* score. *IDFi* (inverse document frequency for word/term i) is given by:

log *(number of documents in collection / number of documents containing word/term i)* +1;

TFij (intradocument term frequency for word/term t in document *j)* is given by:

log *(frequency of word/term* i *in document j)* +1.

The weight of a word/term s in document *j* is therefore *IDFi * TFij.*

Performance of systems was determined by calculating vector cosine measurements between documents. Each system's set of index terms for a document is regarded as a vector in n-dimensional space, where n is the number of words/terms in the entire collection. For SWORD, each vector dimension is given by an individual word; for CLARIT, each dimension is given by phrase or word from a phrase. In this experiment, every CLARIT indexing term was supplemented with the individual words that it contained. Each word received its CLARIT distribution score. When duplicate words were introduced, CLARIT scores were summed. Cosine distance was determined by the formula:

$$COSINE(DOG,DOCj) = ZUi(TERM_{ik} * TERM_{jk}) / \sqrt{\text{EUi } TERM?_k * \text{£l=i } TERMj_k}.$$

A cosine value of 1 indicates complete similarity for two vectors; a value of 0 indicates no similarity.

All processing was fully automatic. In CLARIT's case, a thesaurus for the *AIDS* domain was automatically generated from approximately 8 megabytes of *AIDS* abstracts, none of which covered the forty-two documents in the test corpus. (The resulting thesaurus contained approximately 2,300 words.) Indexing was based on standard CLARIT processing: candidate NPs are identified in the texts, then scored for relevance, matched against the thesaurus, and rescored. (Cf. Evans et al. 1991.) Resulting index terms are derived in part from the thesaurus and in part from the 'novel' terminology in the documents.

3 Results

Results are given in the following tables. In Table 1, the six pairs of 'similar' documents are listed first; the two pairs of 'dissimilar' documents are listed second. SWORD and CLARIT performance is given for each pair of documents and is averaged for each set. In addition, the relative rank of

Full Text	SWORD	CLARIT	CLARIT Improvement
15-16	0.537 <2,2>	0.817 <2,2>	52.1%
20-38	0.386 <2,2>	0.627 <2,2>	62.4%
23-24	0.382 <2,2>	0.702 <3,3>	83.8%
25-26	0.423 <2,2>	0.788 <2,2>	86.3%
27-28	0.385 <3,2>	0.627 <2,2>	62.9%
34-35	0.386 <2,2>	0.761 <2,2>	97.2%
Average:	0.417	0.720	74.1%
4-15	0.197 <11,13>	0.060 <24,41>	228%
36-38	0.208 <13,10>	0.083 <10,40>	151%
"Average;	0.203	0.072	182%

Table 1: Comparison of SWORD vs. CLARIT 'Similarity' Measures on Full Text

Full Text	SWORD "Next-Ranked"	CLARIT "Next-Ranked"
15	0.432 <29>	0.715 <29>
16	0.390 <29>	0.623 <29>
20	0.317 <28>	0.404 <28>
38	0.290 <21>	0.465 <27>
23	0.350 <37>	0.753* <37>
24	0.333 <16>	0.702* <37>
25	0.297 <8>	0.612 <12>
26	0.351 <8>	0.665 <8>
27	0.402* <21>	0.465 <23>
28	0.317 <20>	0.404 <20>
34	0.303 <12>	0.667 <12>
35	0.318 <11>	0.664 <5>
Average:	0.342	0.595
'Difference':	0.075	0.125

Table 2: Comparison of SWORD vs. CLARIT Discrimination on 'Next-Ranked' Documents

each document (as 'seen' by the other) in vector space is indicated in angle brackets. CLARIT's percentage of improvement ('similars' closer to 1, 'dissimilars' closer to 0) is given for each case. Since CLARIT identifies the 'similars' as 'more similar' and the 'dissimilars' as 'less similar' and, further, establishes a better relative ranking of documents than SWORD, we can conclude that CLARIT indexing results in a better model of the document space: documents are 'spread' over a wider range; 'similars' are closer to one another; 'dissimilars' are farther apart. In Table 2, the 'next-ranking' closest document is given for each target document among the 'similars'. This affords a measure of relative discrimination among 'similar' documents. The results indicate that CLARIT establishes relatively significant distances between the 'similar' pairs and the next closest documents.

Additional sample results are given in the appendices. Appendix B gives a portion of a CLARIT index vector; Appendix C, a SWORD index vector. Appendix D contrasts the results for measurements based on document 15. Appendix E repeats and supplements Table 1, including results for just the *abstracts* portions of the forty-two documents.

Such results suggest that basic CLARIT processing is indeed making fine discriminations in the content of documents. Given the apparent properties of the index-term space, we expect similarity measures in very-large-scale databases to support accurate clustering and partitioning of document sets.

4 References

Evans, D.A., Ginther-Webster, K., Hart, M., Lefferts, R.G., Monarch, I.A., "Automatic Indexing Using Selective NLP and First-Order Thesauri." *RIAO '91*, April 2–5, 1991, Autonoma University of Barcelona, Barcelona, Spain, 624–644.

Hersh, W.R., Evans, D.A., Monarch, I.A., Lefferts, R.G., Handerson, S.K., Gorman, P.N. "Indexing Effectiveness of Linguistic and Non-Linguistic Approaches to Automated Indexing." September 1991. In submission to *Medinfo 92*.

A Target Articles

Six similar pairs of articles:

1. **(15) LAMBERT JS, SEIDLIN M, REICHMAN RC, et al, 2,3-Dideoxyinosine (ddl) In Patients With The Acquired Immunodeficiency Syndrome Or AIDS-Related Complex A Phase I Trial, N Engl J Med 1990; 322:1353-40.**
   (16) COOLEY TP, KUNCHES LM, SAUNDERS CA, et al, Once-Daily Administration Of 2,3-Dideoxyinosine (ddl) In Patients With TheAcquired Immunodeficiency Syndrome Or AIDS-Related Complex: Results Of A Phase I Trial, N Engl J Med 1990; 322:1340-5.
2. **(20) FISCHL MA, PARKER CB, PETTINELLI C, et al, A Randomized Controlled Trial Of A Reduced Daily Dose Of Zidovudine In Patients With The Acquired Immunodeficiency Syndrome, N Engl J Med 1990; 323:1009-14.**
   (38) COLLIER AC, BOZZETTE S, COOMBS RW, et al, A Pilot Study Of Low-dose Zidovudine In Human Immunodeficiency Virus Infection, N Engl J Med 1990; 323:1015-21.
3. **(23) BUSCH MB, TAYLOR PE, LENES BA, et al, Screening Of Selected Male Blood Donors For p24 Antigen Of Human Immunodeficiency Virus Type 1, N Engl J Med 1990; 323: 1308-12.**
   (24) ALTER HJ, EPSTEIN JS, SWENSON SG, et al, Prevalence Of Human Immunodeficiency Virus Type 1 p24 Antigen In U.S. Blood Donors - An Assessment Of The Efficacy Of Testing In Donor Screening, N Engl J xMed 1990; 323:1312-7.
4. **(25) GAGNON S, BOOTA AM, FISCHL MA, et al, Corticosteroids As Adjunctive Therapy For Severe Pneumocystis Carinii Pneumonia In The Acquired Immunodeficiency Syndrome A Double-Blind, Placebo-Controlled Trial, N Engl J Med 1990; 323:1444-50.**
   (26) BOZZETTE SA, SATTLER FR, CHIU J, et al, A Controlled Trial Of Early Adjunctive Treatment With Corticosteroids For Pneumocystis Carinii Pneumonia In The Acquired Immunodeficiency Syndrome, N Engl J Med 1990; 323:1451-7.
5. **(27) CLARK SJ, SAAG MS, DECKER WD, et al, High Titers Of Cytopathic Virus In Plasma Of Patients With Symptomatic Primary HIV-1 Infection, N Engl J Med 1991; 324:954-60.**
   (28) DAAR ES, MOUDGIL T, MEYER RD, et al, Transient High Levels Of Viremia In Patients With Primary Human Immunodeficiency Virus Type 1 Infection, N Engl J Med 1991; 324:961-4.
6. **(34) HIRSCHEL B, LAZZARJN A, CHOPARD P, et al, A Controlled Study Of Inhaled Pentamidine For Primary Prevention Of Pneumocystis Carinii Pneumonia, N Engl J Med 1991; 324:1079-83.**
   (35) LEOUNG GS, FEIGAL DW, MONTGOMERY, AB, et al, Aerosolized Pentamidine For Prophylaxis Against Pneumocystis Carinii Pneumonia The San Francisco Community Prophylaxis Trial, N Engl J Med 1990; 323:769-75.

Two non-similar pairs of articles:

1. **(4) BALFOUR HH, CnACE BA, STAPLETON JT, et al, A Randomized, Placebo-Controlled Trial Of Oral Acyclovir For The Prevention Of Cytomegalovirus Disease In Recipients Of Renal Allografts, N Engl J Med 1989; 320:1381-7.**
   (15) LAMBERT JS, SEIDLIN M, REICHMAN RC, et al, 2,3-Dideoxyinosine (ddl) In Patients With The Acquired Immunodeficiency Syndrome Or AIDS-Related Complex A Phase I Trial, N Engl J Med 1990; 322:1333-40.
2. **(36) SCHMIDT GM, HORAK DA, NILAND JA, et al, A Randomized, Controlled Trial Of Prophylactic Ganciclovir For Cytomegalovirus Pulmonary Infection In Recipients Of AUogeneic Bone Marrow Transplants, N Engl J Med 1991; 324:1005-11.**
   (38) HIRSCHEL B, LAZZARIN A, CHOPARD P, et al, A Controlled Study Of Inhaled Pentamidine For Primary Prevention Of Pneumocystis Carinii Pneumonia, N Engl J Med 1991; 324:1079-83.

B Sample CLARIT "Vector"

Article Number: 15

8.6232 IMMUNOGLOBULIN >
8.5745 IMMUNODEFICIENCY >
7.5659 HEMATOLOGIC TOXICITY ?
7.4718 INTRAVENOUS IMMUNOGLOBULIN ?
7.1995 LYMPHOCYTE =
6.8968 P24 =
6.7415 CD-4 =
6.6865 SUBCLINICAL PANCREATITIS ?
5.9430 HEMATOPOIETIC TOXICITY ?
5.9094 LYMPHOCYTE SUBSET ?
5.7093 ZIDOVUDINE =
5.6646 IMMUNODEFICIENCY SYNDROME >
5.5971 PNEUMOCYSTIS >
5.4905 DDI DOSE ?
5.4267 PANCREATIC DYSFUNCTION ?
5.3139 PNEUMOCYSTIS CARINII >
5.2610 KAPOSI >
5.2222 TOXOPLASMA >
5.2140 DDI THERAPY ?
5.1441 RECEIVING DDI ?
5.0872 COMPARE DDI ?
5.0670 DDI CONTENT ?
5.0636 DDI ADMINISTRATION ?
5.0565 DDI TREATMENT ?
5.0306 CARINII =
4.8942 KAPOSI SARCOMA =
4.8392 PANCREATIC LESION ?
4.8117 DEMENTIA >
4.7612 TOXOPLASMA GONDII ?
4.7517 PURINE ANALOGUE ?
4.7076 CD-4 LYMPHOCYTE COUNT ?
4.7020 CLINICAL PANCREATITIS ?
4.6671 ACUTE PANCREATITIS =
4.6247 ANTIVIRAL >
4.6194 SUBSET >
4.6157 ASSOCIATED PANCREATITIS ?
...

Article Number: 16

9.8852 IMMUNODEFICIENCY >
9.6479 LYMPHOCYTE >
8.7882 CD-4 LYMPHOCYTE ?
7.9286 CD-4 =
7.5148 INTRAVENOUS DDI ?
7.4429 P24 =
6.2518 RECEIVED DDI ?
5.9239 CD-4 LYMPHOCYTE COUNT ?
5.7718 IMMUNODEFICIENCY SYNDROME >
5.7093 ZIDOVUDINE =
5.6469 IMMUNODEFICIENCY CLINIC ?
5.2610 KAPOSI >
5.1793 INTRAVENOUS DDI THERAPY ?
5.0306 CARINII >
4.9998 ACQUIRED IMMUNODEFICIENCY >
4.9746 GRANULOCYTE >
4.9702 DDI THERAPY ?
4.9550 HUMAN IMMUNODEFICIENCY >
4.9215 LYMPHOCYTE COUNT >
4.9119 MULTICENTER >
4.8942 KAPOSI SARCOMA =
4.8934 T LYMPHOCYTE >
4.8734 HEMATOLOGIC INTOLERANCE ?
4.8210 ORAL DDI ?
4.7932 PREEXISTING NEUROPATHY ?
4.7387 DDI TREATMENT ?
4.6806 PNEUMOCYSTIS CARINII >
4.6674 LYMPHOMA >
4.5984 VIRAL REPLICATION ?
4.5446 ABNORMALITY =
4.4742 CD8 LYMPHOCYTE COUNT ?
4.4741 HUMAN IMMUNODEFICIENCY VIRUS P24 >
4.4523 CD-4 LYMPHOCYTE COUNT BELOW ?
4.4206 REPLICATION =
4.3487 ASSAY >
4.3305 PNEUMOCYSTIS >
...

C Sample SWORD "Vector"

Article Number: 15	Article Number: 16
19.493371 DDI	19.099575 DDI
13.859593 PANCREATITI	12.681775 NEUROPATHI
12.454870 NEUROPATHIC	10.001659 SUSTAIN
12.018296 NEUROPATHI	9.942532 TRANSAMINAS
9.942532 TP	9.651421 PAINFUL
9.942532 PANCREATIC	9.495894 2'
9.942532 MYER	9.356234 3'
9.942532 METABOLIC	8.745206 PANCREATITI
9.942532 DDA	8.683862 DIDEOXYNUCLEOSID
9.942532 BRISTOL	8.224273 F
9.412165 KILOGRAM	8.197934 TOXIC
9.356234 RECHALLENG	8.021572 CONSTIPATION
9.056558 MAXIMAL	7.997368 CARDIAC
8.745206 AMYLAS	7.745638 INCREA
8.683862 URIC	7.636970 CONDUCTION
8.487884 PURIN	7.636375 DOSE
8.487884 PALNFUL	7.464882 ANALOGU
8.487884 AMINOTRANSFERA	7.439941 KILOGRAM
8.274881 DOSE	7.421133 GRAD
8.162926 PHARMACOKINETIC	7.092311 MAXIMAL
8.021572 YARCHOAN	7.033237 RESISTANT
8.021572 SQUIBB	7.033237 DIDEOXYINOSIN
8.021572 MOIETI	7.033237 BOSTON
8.021572 DIDEOXYCYTIDIN	6.892765 ELEVATION
8.021572 ACCUMULATION	6.847972 URAT
7.687170 GAIN	6.847972 MYELOSUPPRESSION
7.636970 KG	6.847972 MERIEUX
7.636970 DIDEOXYNUCLEOSID	6.847972 KINAS
7.636970 2'	6.847972 HYPERURICEMIA
7.464882 ANALOGU	6.847972 HYDROXID
7.421133 PAIN	6.847972 DIDEOXYADENOSIN
7.421133 GRAD	6.847972 CREATIN
7.202589 TOXIC	6.847972 CHEMOTHERAPEUTIC
7.184730 TOXICITI	6.847972 5'
7.184730 TOLERAT	6.709735 CUBIC
7.033237 EXTREMITI	6.579481 ZIDOVUDIN
7.033237 DIDEOXYINOSIN	6.564945 HEPATIC
7.033237 3'	6.532087 NERVOU
...	...

D Sample CLARIT and SWORD Results

CLARIT Results for Document 15			Sword Results for Document 15		
Doc	Cosine	Rank	Doc	Cosine	Rank
15	+1.0000000000000000e+00	1	15	+1.0000000000000000e+00	1
16	+8.1684521548215350e-01	2	16	+5.3744806083198837e-01	2
29	+7.1456947526138714e-01	3	29	+4.3235081299848827e-01	3
9	+4.4658313382598158e-01	4	23	+3.1387065059682789e-01	4
23	+4.4248119329173685e-01	5	37	+2.9145275580112756e-01	5
24	+4.3595985755939975e-01	6	24	+2.9106913854879102e-01	6
12	+4.0710316300260474e-01	7	21	+2.5472466603145322e-01	7
13	+4.0212540763628962e-01	8	17	+2.2041076979652507e-01	8
37	+3.8284266805599843e-01	9	20	+2.1164956008435837e-01	9
34	+3.5392412347405933e-01	10	27	+2.0619017678072912e-01	10
11	+3.5158517849763316e-01	11	36	+2.0056897705055154e-01	11
18	+3.3167393808030332e-01	12	34	+2.0022596368063181e-01	12
33	+3.2661893117868840e-01	13	4	+1.9743627448006312e-01	13
40	+3.1551692127926101e-01	14	31	+1.9404924805459467e-01	14
35	+3.1021714007446594e-01	15	38	+1.8901007554615129e-01	15
5	+3.0662842722914146e-01	16	28	+1.8276694212913785e-01	16
14	+3.0004512974929221e-01	17	30	+1.8251542579150007e-01	17
8	+2.8504126124539103e-01	18	35	+1.8130327010821246e-01	18
39	+2.7378133890507589e-01	19	7	+1.7978012991447218e-01	19
6	+2.6838490787766933e-01	20	11	+1.7780909714106402e-01	20
27	+2.6542815325638969e-01	21	12	+1.7079485712173351e-01	21
41	+2.6511317715879673e-01	22	9	+1.6928710815012687e-01	22
10	+2.3901705933360240e-01	23	18	+1.6893361581827795e-01	23
19	+2.3808139007039691e-01	24	13	+1.6778288013778114e-01	24
31	+2.2371562607316817e-01	25	32	+1.6265248438975627e-01	25
38	+2.2303213330535573e-01	26	8	+1.5332085195634976e-01	26
42	+2.1529303354773693e-01	27	10	+1.4983204662373417e-01	27
25	+2.1305923907706123e-01	28	26	+1.4364480741613947e-01	28
30	+2.1205467101273509e-01	29	5	+1.4305225468010438e-01	29
20	+2.0816767244943865e-01	30	40	+1.4109811570770067e-01	30
22	+2.0789216369313793e-01	31	33	+1.4042226989740528e-01	31
26	+1.9581858736153696e-01	32	22	+1.4007499721589384e-01	32
1	+1.8797376443988767e-01	33	41	+1.3897763288462459e-01	33
28	+1.6576500005532221e-01	34	2	+1.3674311200022399e-01	34
32	+1.0783947681060313e-01	35	14	+1.3149136027468686e-01	35
i	+9.8911036262129257e-02	36	39	+1.3094755925291698e-01	36
21	+9.6771691187450645e-02	37	25	+1.2884342128806689e-01	37
3	+9.2703770971328092e-02	38	19	+1.2205827056618622e-01	38
17	+6.2760398710365081e-02	39	3	+1.0628959975766347e-01	39
2	+6.2094115526491710e-02	40	6	+1.0283347236708272e-01	40
4	+6.0076167848899589e-02	41	1	+8.0853806787537910e-02	41
36	+4.7215927406381836e-02	42	42	+3.6455426043306507e-116	42

E Results for Full Text and Abstracts

Full Text	SWORD	CLARIT-T	CLARIT-T+N
15-16	0.537 <2,2>	0.873 <2,2>	0.817 <2,2>
20-38	0.386 <2,2>	0.822 <2,2>	0.627 <2,2>
23-24	0.382 <2,2>	0.777 <2,2>	0.702 <3,3>
25-26	0.423 <2,2>	0.708 <2,2>	0.788 <2,2>
27-28	0.385 <3,2>	0.623 <2,2>	0.627 <2,2>
34-35	0.386 <2,2>	0.825 <2,2>	0.761 <2,2>
Average:	**0.417**	**0.771**	**0.720**
4-15	0.197 < 11,13:>	0.166	0.060 <24,41>
36-38	0.208 < 13,10:>	0.199	0.083 <10,40>
Average:	**0.203**	**0.183**	**0.072**

Abstracts	SWORD	CLARIT-T	CLARIT-T+N
15-16	0.562 <2,2>	0.817 <2,3>	0.703 <2,4>
20-38	0.336 <2,2>	0.549 <2,4>	0.509 <2,2>
23-24	0.311 <2,2>	0.688 <4,9>	0.663 <3,3>
25-26	0.343 <2,2>	0.735 <2,2>	0.646 <2,2>
27-28	0.354 <2,2>	0.680 <4,3>	0.653 <2,2>
34-35	0.365 <2,2>	0.504 <3,15>	0.691 <2,3>
Average:	**0.379**	**0.662**	**0.644**
4-15	0.108 < 14,18:>	0.000 < 40,42 >	0.000 <38,42>
36-38	0.119 <10,17:>	0.013 <13,42>	0.007 <20,42>
Average:	**0.114**	**0.007**	**0.004**