Text Analysis Integration into a Medical Information Retrieval System: Challenges Related to Word Sense Disambiguation

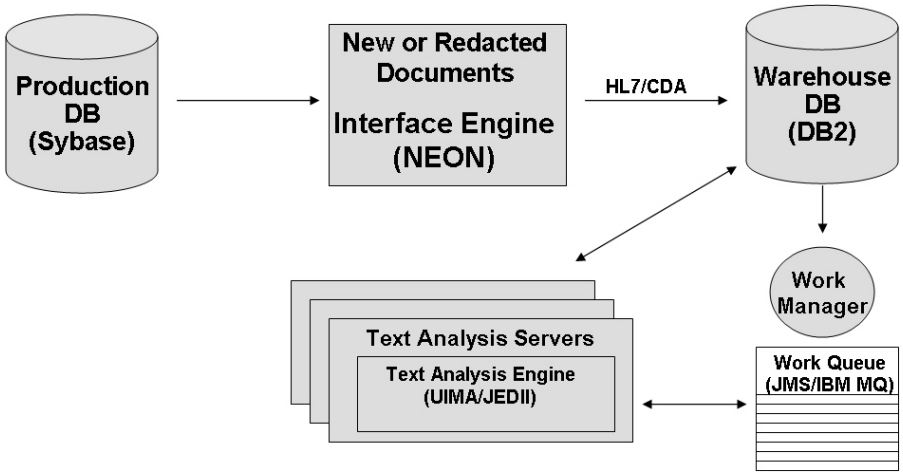
A.R. Coden PhD[1], G.K. Savova PhD.[2], J.D. Buntrock MS[2], I.L. Sominsky MS[1], P.V. Ogren MS[2], C.G. Chute MD, DrPH[2], P.C. de Groen MD[2]

[1] IBM T.J. Watson Research Center, Hawthorne, New York

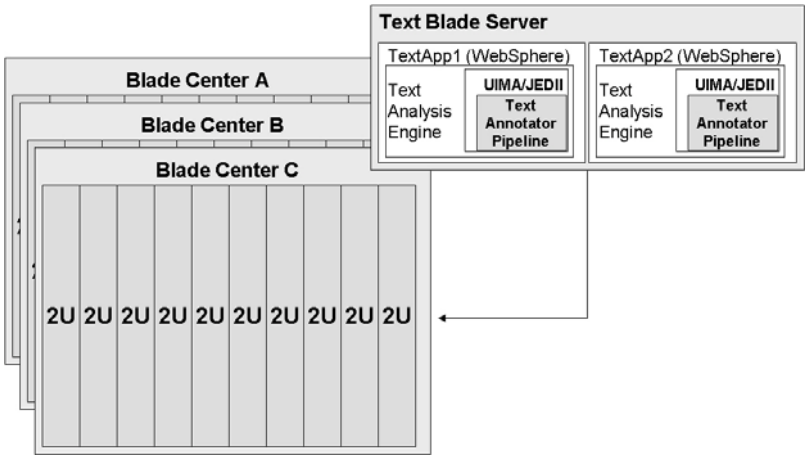
[2] Division of Biomedical Informatics, Mayo Clinic, Rochester, Minnesota

Text analysis production system

- **Large scale, highly modularized, real-time-enabled system**

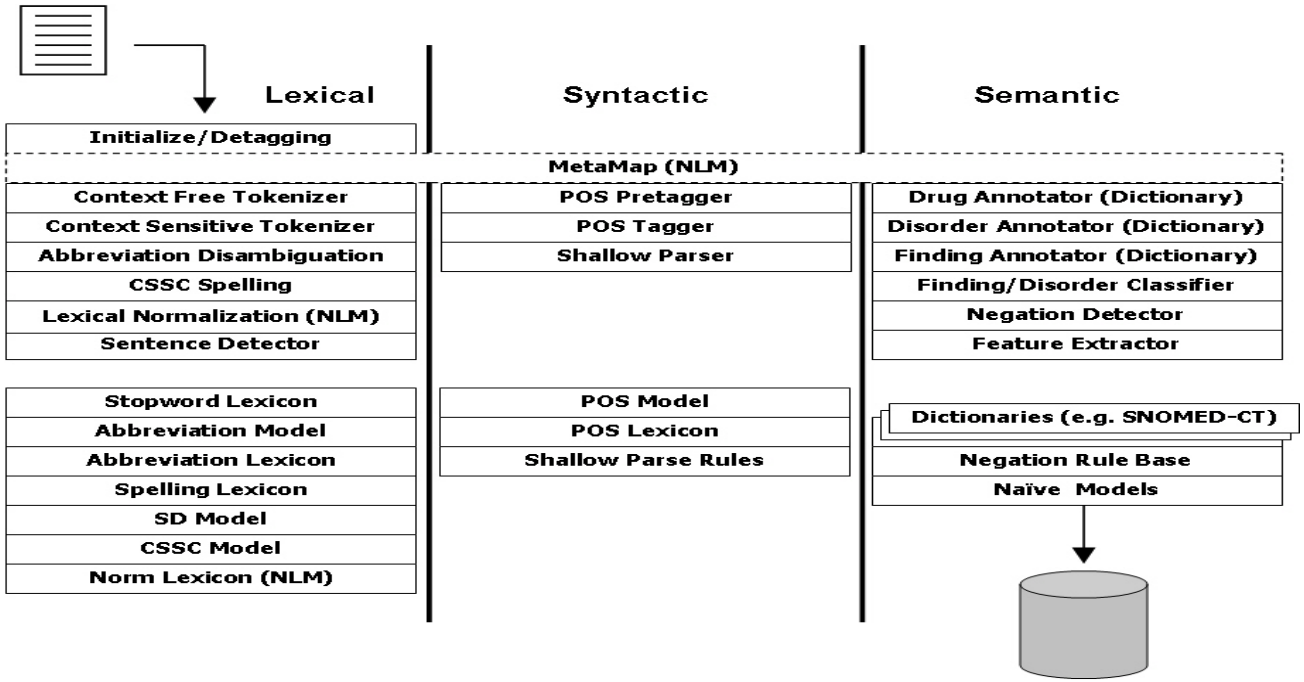


Text Analysis Process Flow



Text Analysis Deployment

Text Analysis System



- Finite state machine – tokenizer, shallow parser, dictionary lookup, negation detector
- Machine learning – abbreviation disambiguator, sentence detector, part-of-speech tagger, finding/disorder/drug classifier

Word Sense Disambiguation Corpora

- Publicly-available corpus based on MedLine abstracts – manually sense-tagged with UMLS concepts
- Built a comparable corpus for clinical ambiguities
  - 50 most relevant clinical terms/phrases
- Comparison
  - 9 terms in common between corpora

Manual Annotation

discharge				
CUIs	Definitions			
C0012621	Discharge, Body Substance			
C0600083	Discharge, Body Substance, Sample			
C0030685	Patient Discharge (Health Care Activity)			
Instance #	**Annotator 1**	**Annotator 2**	**Consensus**	**Text**
1	C0030685	C0030685	C0030685	Patient did not get to school after being **discharged** from the hospital due to there being multiple illnesses going around in the community.
2	C0012621	C0012621	C0012621	She and her family are instructed on signs and symptoms of wound infection including but not limited to pain, erythema, **discharge**, fever or chills.

- Interannotator agreement
  - Number of instances where the annotators agreed over all instances
  - Average: 0.772

WSD algorithm

- Built model for each ambiguity
  - Modification of Huber's robust loss for regression algorithm
- Experimented with 28 different feature sets
- Averaged majority sense: 0.6032
- Average F-score: 0.776
  - Feature set: bag of unique stems, filtered with a tf*idf formula, windows size=10, nonstop list of tokens

Clinical WSD corpus

Ambiguity	Number of Instances	Number of Senses	Sense 1 (majority sense)	Sense2	Sense3	Sense4	Sense5	Sense6	Sense7	Sense8	Sense9	Sense10	Sense11	Sense12	None (included in sense1-sense12)	IAA	F-score
ac	100	12	**59**	9	7	5	5	4	3	2	2	2	1	1	0	0.590	0.776
adjustment	100	5	**82**	12	4	1	1								1	0.740	0.859
affect	100	2	**50**	50											0	0.940	0.965
aid	100	5	**47**	39	9	4	1								9	0.490	0.714
apc	100	6	**84**	6	3	3	3	1							3	0.910	0.844
aspiration	100	2	**59**	41											0	0.990	0.910
block	100	9	**34**	19	13	13	12	3	3	2	1				12	0.220	0.623
burn	100	6	**82**	8	7	1	1	1							1	1.000	0.832
cat	100	3	**50**	48	2										0	0.970	0.960
cervical	100	4	**43**	41	14	2									0	0.900	0.869
cf	100	8	**62**	17	9	4	4	2	1	1					1	0.760	0.914
cold	100	6	**64**	14	14	4	3	1							1	0.920	0.643
compression	100	4	**56**	38	4	2									0	0.920	0.760
condition	100	5	**84**	9	5	1	1								0	0.740	0.859
dilatation	100	2	**52**	48											0	0.970	0.840
discharge	100	2	**65**	35											0	0.970	0.860
drain	100	4	**47**	29	23	1									0	0.820	0.573
dress	100	6	**35**	24	23	11	6	1							0	0.420	0.653
drink	100	4	**54**	42	3	1									0	0.370	0.563
fast	100	4	**85**	13	1	1									0	0.970	0.859
fistula	100	3	**78**	20	2										0	0.750	0.800
fit	100	5	**68**	13	11	6	2								6	0.850	0.710
glass	100	4	**55**	38	5	2									0	0.430	0.910
grade	100	6	**38**	31	17	9	4	1							0	0.520	0.764
interaction	100	5	**72**	20	4	3	1								4	0.910	0.784
iron	100	4	**70**	24	3	3									0	0.900	0.794
irritate	100	3	**84**	15	1										0	0.760	0.864
iv	100	4	**45**	23	20	12									0	0.810	0.590
lead	100	5	**47**	26	17	7	3								7	0.600	0.623
lift	100	7	**87**	5	2	2	2	1	1						0	0.590	0.909
ms	1000	10	**93.1**	47	6	5	3	2	2	2	1	1			2	0.978	0.933
pa	100	7	**56**	34	4	2	2	1	1						2	0.900	0.883
pack	100	5	**79**	15	3	2	1								3	0.140	0.874
patch	100	6	**73**	18	5	2	1	1							0	0.750	0.869
plaque	100	3	**49**	39	12										0	0.480	0.970
ra	100	9	**40**	23	20	6	5	3	1	1	1				5	0.710	0.837
relative	100	2	**53**	47											0	0.650	0.690
sense	100	4	**43**	26	18	13									0	0.460	0.810
sensitivity	100	7	**39**	22	21	12	3	2	1						3	0.450	0.684
sob	1000	2	**98.3**	17											0	0.990	0.983
splint	100	4	**70**	28	1	1									1	0.950	0.768
spot	100	8	**39**	20	14	11	7	5	3	1					20	0.700	0.513
stage	100	10	**28**	17	16	12	7	6	5	4	3	2			5	0.080	0.470
strain	100	9	**35**	18	17	10	7	6	5	1	1				18	0.570	0.535
stress	100	3	**77**	22	1										22	0.980	0.794
support	100	6	**38**	25	13	11	9	4							11	0.620	0.530
tear	100	7	**59**	17	11	6	3	3	1						3	0.900	0.653
transfer	100	2	**84**	16											0	0.840	0.980
valve	100	3	**52**	43	5										0	1.000	0.640
vesicle	100	3	**73**	25	2										0	0.220	0.760
average			**60.328**													**0.722**	**0.776**

Challenges

- Terms have many senses distributed thinly over 100 instances
  - Example: ac
- Terms have high representation of one specific sense
  - Example: ms
- “None of the above category”
  - Does not represent a monolithic sense

Comparison to NLM corpus

Ambiguity	*Number of senses in Mayo WSD set (clinical notes)*	*Number of senses in NLM WSD set (literature abstracts)*	*Best F-score on NLM WSD set: Liu, Teller and Friedman*	*Best F-score on NLM WSD set: Leroy and Rindflesch*	*Best F-score on Mayo WSD set: current study*
adjustment	5	4	not included in study	0.620	0.930
cold	6	6	0.909	not included in study	0.720
condition	6	3	not included in study	not included in study	0.900
discharge	2	3	0.908	not included in study	0.950
fit	5	3	not included in study	not included in study	0.740
lead	5	3	0.910	not included in study	0.900
sensitivity	7	4	not included in study	0.700	0.780
strain	9	3	not included in study	not included in study	0.690
support	6	3	not included in study	not included in study	0.740

Comparison to NLM corpus

- Senses do not necessarily overlap for the 9 terms between the corpora
- In general, more senses per term in clinical notes than in literature abstracts
- For best F-scores on clinical corpus, distinct feature sets for each term

Conclusion

- Built medical information retrieval system
  - Clinical notes accessible via keyword and concept search
  - Near real-time processing
  - Facilitates identification of patient cohorts
  - Nessary building block for other NLP applications
- Ambiguity of terms: source of errors
  - Create a manually sense-tagged set of 50 frequent and highly relevant terms in the clinical domain
  - Integrated a high accuracy sense disambiguator into the medical information retrieval system

References

[1] Friedman C. Towards a Comprehensive Medical Language Processing System: Methods and Issues.
Paper presented at: American Medical Informatics Association (AMIA), 1997.
[2] Aronson AR, Bodenreider O, Chang HF, et al. The NLM Indexing Initiative. Proc AMIA Symp. 2000:17-21.
[3] Frank E, Hall M, Trigg L, Holmes G, and Witten IH. Bioinformatics Advanced Access. Bioinformatics online. April 8, 2004 2004.
[4] Goldin IM, Chapman WW. Learning to Detect Negation with 'Not' in Medical Texts. Paper presented at:
Workshop on Text Analysis and Search for Bioinformatics at the 26th Annual International ACM SIGIR Conference, 2003.
[5] Mack R, Mukherjaa S, Soffer A, et al. Text analytics for life science using the Unstructured Information Management Architecture.
IBM Systems Journal. 2004;48(3):490.
[6] Ferrucci D, Lally A. UIMA: an architectural approach to unstructured information processing in the corporate research environment.
Natural Language Engineering. 2004;10(3:4):327-348.
[7] Liu H, Teller V and Friedman C. A Multi-aspect Comparison Study of Supervised Word Sense Disambiguation.
JAMIA, 2004, vol. 11, num 4, July/Aug.
[8] Leroy G and Rindflesch T. Using Symbolic Knowledge in the UMLS to Disambiguate Words in Small Datasets with a Naïve Bayes Classifier.
Proc Medinfo.2004
[9] Weeber M, Mork JG, Aronson AR. Developing a test collection for biomedical word sense disambiguation. Proc AMIA Symp. 2001.
[10] Thompson-McInness B, Pakhomov S, Pedersen T. Automating Spelling Correction Tools Using Bigram Statistics.
Paper presented at: Medinfo Symposium, 2004; San Francisco, CA, USA.
[11] Pakhomov S. Semi-Supervised Maximum Entropy Based Approach to Acronym and Abbreviation Normalization in Medical Texts.
Paper presented at: 40th Meeting of the Association for Computational Linguistics (ACL 2002), 2002; Philadelphia, PA.
[12] Zhang T and Johnson DE. A Robust Risk Minimization Based Named Entity Recognition System. Proc CoNLL 03, pp. 204-207. 2003.
[13] Coden A, Pakhomov S, Ando R, Chute C. Domain-specific language models and lexicons for tagging. JBI. 2005
[14] Pakhomov S, Coden A, Chute C. Developing a Corpus of Clinical Notes Manually Annotated for Part-of-Speech.
IJMI (Special Issue on Natural Language Processing in Biomedical Applications) 2005.
[15] Chapman WW, Bridewell W, Hanbury P, Cooper G, Buchanan B. Evaluation of Negation Phrases in Narrative Clinical Reports.
Proc AMIA, 2001. Washington, DC, USA.
[16] Zhang T. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In ICML'04, pp. 919–926.

Acknowledgements: We would like to thank Debra Albrecht, Barbara Abbott, Pauline Funk, Donna Ihrke for their professionalism and meticulous work in manually tagging the data.
Address for correspondence: anni@us.ibm.com