HHS Public Access

Author manuscript

Proc ACM Int Workshop Data Text Min Biomed Inform. Author manuscript; available in PMC 2016 April 11.

Published in final edited form as:

Proc ACM Int Workshop Data Text Min Biomed Inform. 2013 ; 2013: 9–12. doi:
10.1145/2512089.2512101.

Document Sublanguage Clustering to Detect Medical Specialty in Cross-institutional Clinical Texts

Kristina Doing-Harris,

Department of Biomedical Informatics, University of Utah, Health Sciences Center, Salt Lake City,
UT

Olga Patterson,

VA SLC Health Care, Salt Lake City, UT

Sean Igo, and

Department of Biomedical Informatics, University of Utah, Health Sciences Center, Salt Lake City,
UT

John Hurdle

Department of Biomedical Informatics, University of Utah, Health Sciences Center, Salt lake City,
UT

Kristina Doing-Harris: kristina.doing-harris@utah.edu; Olga Patterson: olga.patterson@utah.edu; Sean Igo:
sean.igo@utah.edu; John Hurdle: John.Hurdle@utah.edu

Abstract

This paper reports on a set of studies designed to identify sublanguages in documents for domain-specific processing across institutions. Psychological evidence indicates that humans use context-specific linguistic information when they read. Natural Language Processing (NLP) pipelines are successful within specific domains (i.e., contexts). To limit the number of domain-specific NLP systems, a natural focus would be on sublanguages. Sublanguages are identified by shared lexical and semantic features.[1] Patterson and Hurdle[2] developed a sublanguage identification system that functioned well for 12 clinical specialties at the University of Utah. The current work compares sublanguages across institutions. Using a clinical NLP pipeline augmented by a new document corpus from the University of Pittsburgh (UPitt), new documents were assigned to clusters based on the minimum cosine-distance to a Utah cluster centroid. The UPitt documents were divided into a nine-group specialty corpus. Across institutions, five of the specialty groups fell within the expected clusters. We find that clustering encounters difficulty due to documents with mixed sublanguages; naming convention differences across institutions; and document types used across specialties. The findings indicate that clinical specialty sublanguages can be identified across institutions.

Keywords

Medical informatics applications; natural language processing; cognitive science

Competing Interests

The authors declare they have no competing interests in this work.

1. INTRODUCTION

We find it useful to use Cognitive Science theories of language understanding to guide NLP system development. One such finding is domain-dependent processing.[3] Linguistics gives an indication of the domains associated with human processing by defining sublanguages. In previous work we found that clustering techniques could identify the domain (clinical specialty) of a provider's note based on sublanguage properties.[2] This study considers whether cluster domains from one institution will remain consistent across others. In particular, we examine clinical documents from the University of Utah with those from the University of Pittsburgh NLP Repository.

1.1 Background

Machine learning researchers joke, that “as long as your sentence appeared in the Wall Street Journal between 1980 and 1990, then we have solved the problem of NLP.” It is well recognized within the NLP community that language models do not generalize well.[4] There seems to be consensus that this failure is due in part to the variability of language across domains. One approach to this problem is to use domain adaptation technology to modify a language model acquired using one domain and optimize it for a different domain. [5] Our approach responds instead by focusing on a set of “solved domains” and then determining to which domain a given new document is most closely related.

1.2 Domain-Specific Processing

We are developing an approach to NLP that is based on findings from cognitive science by incorporating the psychological evidence for context-specific language processing.[6] The context of a piece of text or spoken language includes the expectations about word sense based on statistical usage.[7] Cognitive studies using eye tracking and event related potentials (ERPs) find that reading times increase and brain electrical patterns change when statistical expectations are violated. These experiments find that statistical expectations vary based on the topic, author, or setting of the text.[7], [8] For example, the expected acronym expansion for “cc” in lab values would be “cubic centimeters”; while for email note interpretation it would be “carbon copy.”

In NLP, the differences in statistical expectations for words and their meanings are referred to as the domain-specific language model. A common approach to developing such models for document search and clustering involves creating profiles based on the distributions of single words and multi-word phrases.[9] These distributions depend on the word and sentence context. Context in NLP refers most often to co-occurring text (i.e., the surrounding words). The profiles and contexts are quite specific to the particular subfield and may evolve with time.

When aiming to improve processing accuracy, NLP developers limit the system focus to a small domain. However, this approach seems infeasible because it may lead to an explosively large number of domain-specific processors. Establishing the boundaries of a domain, which are narrow enough to provide good NLP results, but broad enough to keep

the number of developed systems manageable, is a step toward optimal computerized language processing.

Linguistics findings provide an insight into possible domain boundary solutions. Human language processing follows different paths based on different word and sense expectations. When a specialized domain imposes restrictions on creating valid sentences and has a relatively narrow vocabulary, it is said to have developed its own sublanguage.[10]

1.3 Clinical Sublanguages

Research efforts show that the language used in clinical settings differs from non-medical language[11] and biomedical literature.[12] It is also recognized that clinical narratives may have multiple sublanguages.[13] Clinical NLP researchers suspect that terminology (medical terms) and semantic language structure change depending on medical specialty, author's clinical role (physician, nurse), and clinical setting (inpatient, outpatient). Of these, Patterson and Hurdle[2] showed that medical specialty sublanguage vocabulary allowed unsupervised document clustering to create very pure specialty specific clusters.

Directing text to one of multiple paths, requires a method for assigning a text to a path. We propose that each path corresponds to a sublanguage, and that identifying the sublanguage would enable appropriate assignment of a document to a processing path.

1.4 Objective

Building on the work of Patterson and Hurdle, we set out to determine if lexical and semantic features used by the same medical specialty across institutions provide enough indication of sublanguage to allow identified by document clustering. Robust sublanguage discrimination meets the following criteria: 1) documents from each institution form pure clusters by clinical specialty; 2) additional documents align by specialty with the original clusters; 3) adding documents from new clinical specialties creates new pure clusters into which documents from corresponding specialties in the original group will cluster.

2. METHODS

For consistency across experiments, feature vector definitions and clustering algorithms from the Patterson and Hurdle study were used (shown in Figure 1). Detailed description of the feature vectors and clustering algorithm are in [2]. Processing was done on a high-performance HIPAA-compliant compute cluster.

2.1 Document Sets

Utah—The Patterson and Hurdle study used 12 note types. They were found to correspond to 12 pure document clusters by clinical specialties. The (Utah) dataset consists of 3000 randomly selected notes from each of 12 document types created from January 2007 to December 2008 at the University of Utah Hospital. The corpus spanned clinician roles, specialties, and environments.

UPitt—The University of Pittsburgh NLP Repository (UPitt) corpus contains 45,364 documents. They were provided in five document type directories (listed in Table 1 along

with the count of subdirectories). Based on a preliminary clustering across all documents, we found that the main document types were gathered from multiple clinical specialties. However, by combining lower-level subdirectories that were specialty specific we created specialty-specific groups of notes.

2.2 Feature Extraction

As described in Patterson and Hurdle [2] and illustrated in Figure 1, our feature vectors rely on a ‘bag of words & semantics’ vector-space model, restricted to only those terms unambiguously mapped by MetaMap [14] to the UMLS Metathesaurus and their semantic types. This corresponds to the domain keyword of the sublanguage. Other sublanguage features were not included. Term-frequency-inverse-document-frequency (TF-IDF) statistics were used to quantify features. Feature vectors were created by:

1. Each file was processed with MetaMap (v. 2009) to identify any phrasal term that could be mapped to at least one concept and any semantic type that could be *unambiguously* mapped. [15]
2. Multiword terms were tokenized by white space and all tokens were normalized using the SPECIALIST lexicon Norm tool (LVG) to decrease the size of the feature space. TF-IDF scores were calculated for the resulting tokens.
3. Words that appeared in over 95% of a particular note type were filtered out because they were presumed to be boilerplate.
4. IDF values were calculated for each term across the Utah 12-specialty corpus. A second common word list and stopwords list, which contained 30 terms (e.g. tmco, jan, progress), removed common words and those that might cause large clusters to form.
5. Final TF-IDF feature vectors were then created.

Despite attempts to minimize the size of the vector space, feature vectors were sparse and still quite large, on the order of 50,000 features.

2.3 Clustering Algorithm

The algorithm used was the CLUTO implementation of bisecting K-means with end-time global optimization (-rbr), including the default settings for cluster selection method (best) and criterion function (I^2). [16] Bisecting K-means is a hierarchical algorithm, which splits the feature space vectors into the predefined number of clusters by iteratively separating them based on cohesion and separation.

2.4 Comparison: UPitt to Utah clusters

To compare the UPitt documents to the documents from Utah, the feature vectors for the UPitt documents were extracted and the feature set was restricted to only those features that also occurred in Utah documents. This removed 300 out of approximately 50,000 features from the UPitt documents. For the IDF values for each feature, we used the 12-specialty Utah IDF values in order to maintain consistency with the Utah document set. These IDF values will also facilitate the future clustering of single documents.

Document assignment to clusters was based on the cluster similarity measurement, which is the minimum cosine distance to the centroid. The centroids for the 12 pure Patterson and Hurdle clusters were calculated as the arithmetic mean of all its feature vectors. The centroid feature vector is a vector whose every coordinate i is the arithmetic mean of coordinate i from all the vectors, which were assigned to that cluster. UPitt documents were assigned to the cluster whose centroid had the smallest cosine distance to the document's feature vector.

3. RESULTS

We determined that there were nine potentially overlapping medical specialties between the Utah and UPitt corpora; Cardiology (CARD), Dermatology (DERM), Family Practice (FAMP), Hematology Oncology (HONC), Neurology (NEUR), Operative Report (OPRP), Orthopedics (ORTH), Plastic Surgery (PLAS), Rheumatology (RHEU).

To determine if new documents fall within appropriate clinical specialty specific clusters, we classified the UPitt documents based on the Utah clusters. The results are reported in Table 2. A dark gray along the diagonal would indicate that UPitt documents align with the Utah clusters. DERM, FAMP, PLAS, and RHEU notes could not be separated as the previous result indicated. DERM were clustered with Utah Hematology Oncology notes, which could be because of the mentions of skin cancer in Utah Hematology Oncology notes. Since PLAS were operative reports and not clinic notes, they shared lexical and semantic features with both plastic surgery clinic and operative reports. RHEU also included Rheumatology Operative Reports, which may share vocabulary with Utah Orthopedic notes. HONC notes, however, clustered well with Utah Hematology Oncology notes.

OPRP and ORTH maintained their secondary groupings. OPRP may have included OB/GYN as there were very few notes in the UPitt OB/GYN directory. We grouped the Orthopedics operative reports from UPitt with the ORTH notes because Ortho at Utah includes orthopedic surgery, which may be why 273 notes clustered with Operative Report. The Utah specialty corpus excluded Utah Family Practice notes because they did not contain a pure sublanguage. The UPitt FAMP clustering with Cardiology is probably reflective of the patient population.

4. DISCUSSION

This work demonstrates that sublanguages are somewhat consistent across institutions based on the clinical milieu of the documents. Clinical setting and clinical subdomain, but not specific institution, influence the determination of unique sublanguages. Separate sublanguages form distinct clusters based on k-means similarity clustering technique.

4.1 Provisos to sublanguage consistency

The uniqueness of the sublanguages in the new corpus carries a caveat. These sublanguages may not fall along the lines of the directory structure used by the institution. An institution may group their documents by type (e.g., History and Physical, Admission note) as opposed to medical specialty. Therefore, it is important to use similarity metric and not the document name to determine sublanguage boundaries.

Note types that are used by multiple clinical specialties or in multiple clinical settings (e.g., Discharge Summary) do not form pure clusters within a single institution nor across institutions. On the other hand, note types that form consistent clusters within one institution fall within the appropriate specialty-specific clusters at other institutions. From the UPitt corpus DERM, HONC, CARD, FAMP and RHEU all clustered together, when compared to the Utah centroids. DERM and HONC fell in the Hematology-Oncology cluster; CARD and FAMP fell in the Cardiology Clinic cluster; and RHEU is a special case because of its very small number of documents. Documents of five note types did not fall in the same cluster, which implies that they belong to separate sublanguages. PLAS from the UPitt corpus clustered with ORTH and OPRP in the single institution clustering experiment as well as with Orthopedics and Plastic Surgery and Operative Reports in the Utah centroid comparison.

4.2 Implications for Document Paths

In the current study we analyzed documents as a whole and utilized document type as a proxy for the clinical subdomain and clinical setting. However, previous research has suggested that multiple sublanguages might be used within a single document.[13] Text-processing systems take advantage of the layout of a document to localize the information of interest.[17] This is particularly true for loosely templated documents, such as a progress note based on the widely used SOAP model.[18]

Our intent is not to ignore these alternate groupings. We see document paths as multiple and overlapping, as suggested by the cognitive studies.[8] In this case, there may be a language-processing path for ER document and a path for the highly structured physician note. A single note may qualify for both paths. The interpretation is then dependent on the first path to produce an output, which best satisfies the constraints across layers. By suggesting a multitude of paths we again risk a combinatorial explosion and a waste of resources by duplicative processing. For instance, our current PGN and OPRP clusters encompass many of the Plastic Surgery and OB/GYN notes. Should we disband PGN and OPRP in favor of more specific specialty note groups? How specialized should those groups get? These are questions we think should be answered pragmatically.

4.3 Future Work

The first round of experiments suggested that there are substantial differences in the language used by different clinical subdomains.[2] The results of the current set of experiments support previous findings across multiple institutions, and indicate that clinical setting also affects language use. The Multi-parameter Intelligent Monitoring in Intensive Care Database provided by the University of Minnesota (MIMIC II) contains a set of notes are related to cardiology and were not included in the current analysis. However, they provide another axis for language analysis -- grouping by author (physicians, nurses, technologists).

Once sublanguage boundaries are clearly identified and described, our next step will be to create a processing pipeline that makes use of multi-path processing. The planned system

will enable a flexible architecture to define multiple sublanguage models and the appropriate processing path for each incoming document.

4.4 Limitations

In related studies that we do not have the space to report here, we extended the number of sublanguage clusters by including the UPitt ER and PGN directories. This study ran into two difficulties. The creation of unique clusters was found to be particularly vulnerable to the known clustering problem, sensitivity to initial conditions.[19] New pure clusters did not form unless the new directories were the first feature vectors to be presented. After they formed no other presentation orders were considered. The second difficulty was that ER documents from the Utah corpus do not substantially fall within the expected cluster (UPitt ER). The largest group was within the Utah Cardiology Clinic Note cluster. It is no surprise that ER notes tend to be spread across specialties. Given the prevalence, critical, and emergent nature of cardiac problems in the US, it is also unsurprising that a large number of ER notes use the sublanguage of Cardiology. Manually inspection confirmed that the medical subject matter of a sample of notes agreed with the clinical specialty of the cluster in which they fell.

We also could not cluster the UPitt documents to compare Utah documents because the UPitt specialty groups were too small.

5. CONCLUSIONS

The results of our research indicate that clinical sublanguages are not institution specific. These findings inform future NLP development, establishing sublanguage boundaries for more efficient system scope definition.

ACKNOWLEDGEMENTS

We gratefully acknowledge the BluLab at the University of Pittsburgh for the UPitt corpus, and the University of Utah Center for High Performance Computing for computer time.

Funding

This work was supported by National Library of Medicine award R01-LM010981.

REFERENCES

1. Harris, ZS. A theory of language and information: A mathematical approach. Oxford and New York: Clarendon Press; 1991.
2. Patterson O, Hurdle JF. Document clustering of clinical narratives: a systematic study of clinical sublanguages. AMIA Annu Symp Proc. 2011; 2011:1099–1107. [PubMed: 22195171]
3. Jindal P, Roth D. Using domain knowledge and domain-inspired discourse model for coreference resolution for clinical narratives. JAMI. 2013 Feb; 20(2):356–362.
4. Smith NA, Martins AFT. Linguistic structure prediction with the sparseptron. XRDS. 2013 Mar. 19(3):44.
5. Daumé H III, Marcu D. Domain Adaptation for Statistical Classifiers. J. Artif. Intell. Res.(JAIR). 2006; 26:101–126.
6. Walenski M, Ullman MT. The science of language. The Linguistic Review. 2005; 22(2):327–346.

7. Farmer, TA.; Fine, AB.; Jaeger, TF. Implicit context-specific learning leads to rapid shifts in syntactic expectations; Proc of the 33rd Annu Meet of the Cogn Science Socy; 2011. p. 2055-2060.
8. Traxler, M. Introduction to Psycholinguistics. Wiley-Blackwell; 2012.
9. Krallinger M, Valencia A, Hirschman L. Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome Biol.* 2008; 9(2):S8. [PubMed: 18834499]
10. Harris ZS. The structure of science information. *J Biomed Inform.* 2002 Aug; 35(4):215–221. [PubMed: 12755516]
11. Campbell DA, Johnson SB. Comparing syntactic complexity in medical and non-medical corpora. *Proc AMIA Symp.* 2001:90–94. [PubMed: 11825160]
12. Friedman C, Kra P, Rzhetsky A. Two biomedical sublanguages: a description based on the theories of Zellig Harris. *J Biomed Inform.* 2002 Aug; 35(4):222–235. [PubMed: 12755517]
13. Zeng QT, Redd D, Divita G, Jarad S, Brandt C. Characterizing Clinical Text and Sublanguage: A Case Study of the VA Clinical Notes. *J Health Med Informat S.* 2011; 3:2.
14. Aronson AR, Lang F-M. An overview of MetaMap: historical perspective and recent advances. *JAMIA.* 2010 May; 17(3):229–236. [PubMed: 20442139]
15. Patterson O, Igo S, Hurdle JF. Automatic acquisition of sublanguage semantic schema: towards the word sense disambiguation of clinical narratives. *AMIA Annu Symp Proc.* 2010; 2010:612. [PubMed: 21347051]
16. Zhao Y, Karypis G. Data clustering in life sciences. *Mol. Biotechnol.* 2005 Sep; 31(1):55–80. [PubMed: 16118415]
17. Johnson SB, Bakken S, Dine D, HYUN S, Mendonca E, Morrison F, Bright T, Van Vleck T, Wrenn J, Stetson P. An Electronic Health Record Based on Structured Narrative. *JAMIA.* 2007 Oct; 15(1):54–64. [PubMed: 17947628]
18. Weed LL. The Problem Oriented Record as a Basic Tool in Medical Education, Patient Care, and Research. *Ann Clin Res.* 1971 Jan.3(3)
19. Duda, RO.; Hart, PE. *Pattern Classification and Scene Analysis.* New York: John Wiley & Sons; 1973.

Figure 1.
Patterson and Hurdle (2011) TF/IDF feature vector pipeline.

Table 1

UPitt corpus directory structure

Document Type	Directory Name	Subdir count
Discharge Summary	DS	67
Emergency Room reports	ER	7
History and Physical	HP	66
Operating Report	OR	47
Progress Note	PGN	24

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2

Determining the cluster a document falls within using the centroid method. Dark grey cell indicates the largest group for each document type. Light grey cell indicates secondary groups.

Utah clusters	FAMP [‡]	CARD	HONC	OPRP	NEUR	ORTH	PLAS [*]	RHEU	DERM
Burn Clinic Notes [‡]	0	1	4	3	3	24	17	0	0
Case Man Discharge [‡]	0	11	3	5	17	4	0	0	0
Social Services Note [‡]	0	5	2	12	3	11	0	0	0
Ob/Gyn Clinic Notes [‡]	2	2	6	357	7	9	1	0	0
Cardiology Clinic	23	986	25	49	45	38	2	0	0
Hematology Oncology	10	73	100	12	44	11	5	3	15
Operative Report	0	33	1	1735	2	273	93	1	0
Neurology Clinic	8	30	21	40	625	72	4	1	0
Ortho & Plastic Surgery	2	4	0	5	56	508	37	5	0
Plastic Surgery Clinic	0	3	1	10	15	6	12	0	0
Rheumatology Clinic	1	3	1	1	0	1	0	3	2
Dermatology Clinic	0	1	0	31	14	0	1	0	0
Total	46	1132	164	2315	831	975	172	13	17

^{*}Plastic Surgery,
[‡]These specialties have no match across corpora.