

Unsupervised Induction of Domain-Specific Semantic Frames from Operative Notes for Surgical Clinical Research

Yan Wang, MS¹, Serguei Pakhomov, PhD^{1,2}, Genevieve B. Melton, MD, MA^{1,3}
¹Institute for Health Informatics, ²College of Pharmacy, ³Department of Surgery,
University of Minnesota, Minneapolis, MN

Abstract

Surgical clinical research often utilizes information within operative notes including data about the supplies, tools, and techniques used for a given procedure. We have previously demonstrated that compared to general English semantic frames, surgical operative notes contain semantic frames with additional or different arguments and that additional verb senses. In this study, we investigated the use of an unsupervised approach for inducing semantic frames for surgical action verbs. Our results demonstrate that semantic frames induced from domain corpus using unsupervised approach cover most of the domain semantic frames created by human experts.

Introduction

Information contained in operative notes such as techniques, supplies (e.g., mesh type, prosthetic), or other surgical parameters (e.g., pneumoperitoneum pressure) used are critical to the conduct of surgical clinical research aimed improving evidence-based medical practice. Semantic frames for surgical action verbs form an indispensable foundation for the creation of operative note information extraction (IE) systems to automatically extract the associated knowledge within the large amounts of operative notes stored in Electronic Health Record (EHR) systems [1-3]. Previous studies show that existing semantic general English resources have limited coverage of surgical actions, and surgical action verbs often involve special semantic arguments [4, 5]. These domain-specific semantic frames can be used to build a robust Semantic Role Labeling (SRL) system to extract surgical information from operative notes. In this study, we propose an unsupervised approach for inducing semantic frames for surgical action verbs from Stanford-typed dependencies as well as syntactic and semantic arguments.

Method

A multinomial Dirichlet Process Mixture Model (DPMM) clustering model[6] was built to induce semantic frames from a corpus of 3,000 laparoscopic cholecystectomy (ICD-9-CM Procedure code 51.23) operative notes. As shown in Figure 1, the 3,000 operative notes were first pre-processed to extract surgical action sentences out from each note and then to generate deep parse trees of all these action sentences. For each top surgical action verb, all sentences with the verb in a deep parse tree were collected. The collected action sentences were then parsed by the Stanford Dependency parser[7] afterwards. Initial semantic frames were constructed for each sentence of an associated surgical action verb with arguments collected from the dependency parse as well as the constituent parse of the sentence. To augment clustering performance, we used Unified Medical Language System (UMLS)[8] semantic types collected by using MetaMap[9] mapping of the head nouns in the Stanford Dependency parse instead of the original head noun in the argument (e.g., “dojb:Proedure” and “dojb:Agent”) for clustering. Clusters with large numbers of instances were examined by human experts to identify domain-specific-frames.

Result

The final clustering model was built with hyper-parameters α and β set to 1.0. For 30 verbs performing clustering on a median of 1,189 sentences (45 - 9462) for verb instance, we found that on average, for each verb the clustering algorithm discovered 2 frames. Compared with average frames discovered in general English (average of 15.2 frames per verb)[6], action verbs in operative notes present fewer usage patterns. We evaluated the sensitivity of our approach for discovering domain specific semantic frames on a manually annotated corpus for the top 30 action verbs. Table 1 shows that the model automatically generated 73 frames while human annotators created a total of 40 frames. The 31 modified frames generated by the model included all the 11 modified frames generated by human annotators. Two of the new frames that were discovered with manual annotation were discovered by the model. Overall, 93% of the modified/new frames generated by human annotators were discovered by the unsupervised model.

Conclusion

This study demonstrates that an unsupervised approach can be used to effectively discover domain-specific semantic frames from un-annotated corpus for top surgical actions from operative notes. Future work will include discovering

domain-specific frames for a broader range of surgical actions performed on a broader set of operative notes and to apply these newly discovered frames in for SRL modeling.

Acknowledgements

The authors would like to thank Fairview Health Services and support from American Surgical Association Foundation and National Institutes of Health (#R01LM011364(GM), #R01GM102282(SP), #8UL1TR000114).

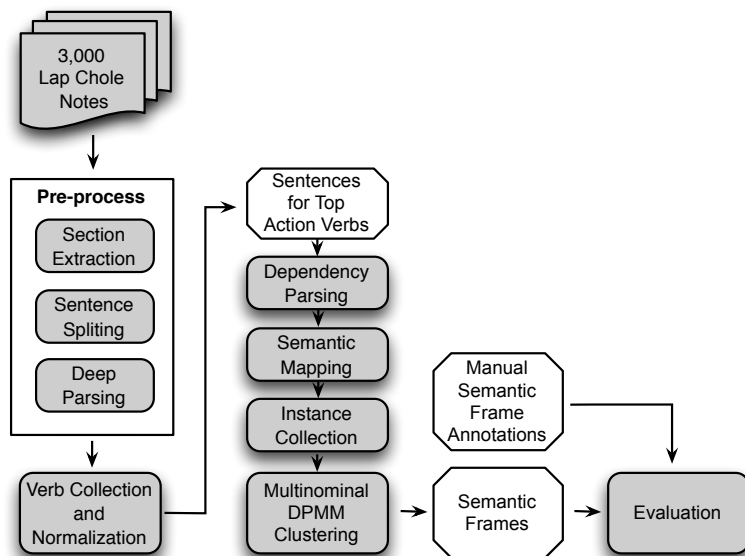


Figure 1. Pre-processing, dependency parsing, instance collection and Multinomial DPMM clustering

Frames Source	Total	Modified Frames	New Frames	New Semantic roles
Human annotated semantic frames	40	11	3	12
Multinomial DPMM induced semantic frames	73	31	2	21

Table 1. Semantic frames induced by Multinomial DPMM clustering vs. human annotated semantic frames

References

- [1] Surdeanu M, Harabagiu S, Williams J, Aarseth P. Using Predicate-Argument Structures for Information Extraction. Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1.
- [2] Wattarujeekrit T, Shah PK, Collier N. PASBio: predicate-argument structures for event extraction in molecular biology. BMC Bioinformatics. [Research Support, Non-U.S. Gov't]. 2004 Oct 19;5:155.
- [3] Cairns BL, Nielsen RD, Masanz JJ, Martin JH, Palmer MS, Ward WH, et al. The MiPACQ Clinical Question Answering System. AMIA Annu Symp Proc. 2011;2011.
- [4] Wang Y, Pakhomov S, Burkart NE, Ryan JO, Melton GB. A study of actions in operative notes. AMIA Annu Symp Proc. [Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't]. 2012;2012:1431-40.
- [5] Wang Y, Pakhomov S, Melton GB. Predicate Argument Structure Frames for Modeling Information in Operative Notes. Stud Health Technol Inform. 2013;192:783 - 7.
- [6] Kawahara D, Peterson D, Popescu O, Palmer M. Inducing Example-based Semantic Frames from a Massive Amount of Verb Uses: Association for Computational Linguistics; 2014.
- [7] Marneffe M-Cd, MacCartney B, Manning CD. Generating Typed Dependency Parses from Phrase Structure Parses. LREC 20062006.
- [8] Unified Medical Language System (UMLS). Available from: <http://www.nlm.nih.gov/research/umls/>.
- [9] Aronson AR, Lang F-M. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc. 2010;17:229-36.