Identifying Key Concepts from EHR Notes Using Domain Adaptation

Jiaping Zheng
College of Information
and Computer Sciences
University of Massachusetts
Amherst, MA
jzheng@cs.umass.edu

Hong Yu
Bedford VA Medical Center
Bedford, MA
Department of Quantitative Health Sciences
University of Massachusetts
Worcester, MA
hong.yu@umassmed.edu

Abstract

Linking electronic health records (EHRs) to relevant education materials can provide patient-centered tailored education which can potentially improve patients' medical knowledge, self-management and clinical outcome. It is shown that EHR query generation using key concept identification improves retrieval of education materials. In this study, we explored domain adaptation approaches to improve key concept identification. Our experiments show that a 20.7% improvement in the F1 measure can be achieved by leveraging data from Wikipedia. Queries generated from the best performing approach achieved a 20.6% and 27.8% improvement over the queries generated from the baseline approach.

1 Introduction

Providing patients with access to their own electronic health records (EHRs) has been shown to benefit patients in many ways, including enhanced medical understanding, and better medication adherence (Delbanco et al., 2012). Several studies have also found that providing knowledge can improve diabetes-related health outcomes (Wiljer et al., 2006).

However, EHR notes present unique challenges to the average patients. A national survey in US shows that 36% of the population have basic or below basic health literacy (National Center for Education Statistics, 2003). The language in the EHR notes is difficult for non-medical professionals to comprehend because of the prevalence of medical terms, abbreviations, and domain-specific language patterns. Coupled with limited average health literacy, the valuable and authoritative information contained in the EHR is less accessible to the patients, who ultimately stand to benefit the most from the information.

Linking EHR notes to relevant education materials can unlock the information in them and provide patient-centered tailored education which has the potential to enhance patient engagement and lead to improved self-management and clinical outcomes. One challenge in designing such an Information Retrieval system is to generate queries. It is shown that ad hoc retrieval using the entire EHR note is less effective because of the noise contained in the notes (Zheng and Yu, 2015). A better strategy is to identify the key concepts from the notes and use them as queries. Using off-the-shelf concept recognition tools such as MetaMap (Aronson, 2001) can lead to long queries that contain many unimportant concepts. For example, incidental findings in an EHR note may distract the retrieval system from returning documents that are central to the note. Therefore, identifying, among all the concepts, the important ones is essential to generate effective queries. In this study, we explored domain adaptation approaches (Jiang and Zhai, 2007; Daumé III, 2007) to improve key concept identification. These approaches have been demonstrated to improve performances of NLP tasks such as semantic role labeling (Dahlmeier and Ng, 2010) and discourse connective detection (Polepalli Ramesh et al., 2012).

Our system in the training phase uses a combination of Wikipedia data and EHR data to learn models to identify key concepts. At the test time, the models are used to predict key concepts from the EHR notes. The identified key concepts are then grouped into one query string to retrieve relevant education documents.

2 Related Work

Domain adaptation is a method to adapt machine learning models trained from a large labeled out-

Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis (Louhi), pages 115–119,
Lisbon, Portugal, 17 September 2015.

of-domain dataset to a target domain in which labeled data is difficult to obtain. Due to privacy regulations, limited health care data is readily available to train machine learning models (Chapman et al., 2011). Thus, domain adaptation approaches are investigated in many NLP tasks. In Part of Speech tagging, Coden et al. (2005) showed combining Penn Treebank data with a small clinical notes corpus improves performance. Liu et al. (2007) developed a heuristic sample selection method to select training samples from the medical domain, and combined with Penn Treebank data to adapt a maximum entropy tagger.

There is also interest in adapting models in other NLP tasks. Polepalli Ramesh et al. (2012) showed that domain adaptation techniques yielded the best performance in identifying discourse connectives in biomedical text. Kim et al. (2013) extracted congestive heart failure related mentions by adapting models learned from a different type of clinical notes.

Information Retrieval in the biomedical domain is also related to this work. The CLEF eHealth (Kelly et al., 2014) challenge includes a task to retrieve information to address questions patients may have when reading clinical reports. This task provides participants with expert-formulated concise queries for one central disorder in discharge summaries (Goeuriot et al., 2014). In our study, we aim to generate queries from long EHR notes without the help of experts. TREC Clinical Decision Support Track is another information retrieval challenge involving EHR notes. The task is designed to address the physicians' information needs rather than the patients'. Case reports are provided as query descriptions, which can be shorter and more focused than an EHR note.

3 Materials

Twenty progress notes are randomly selected from a de-identified corpus of EHR notes to test our systems' performance. Each note contains on average 261 tokens, with a standard deviation of 133. A physician read each note, and manually identified relevant education materials from a collection of MedlinePlus[1] documents. The phrases in the EHR notes that match the title of a relevant MedlinePlus document are marked as key concepts. A snippet of one note with its linked education materials is show in Table 1. Key concepts marked by matching titles are *italicized*.

Snippet of EHR
Patient is a XX-year-old woman status post Thoratec left ventricular assist device placement for *cardiogenic shock* following acute myocardial infarction. Patient requires critical care for management of her *respiratory failure*, malnutrition, hyperglycemia, post-procedure hemodynamics, and *renal failure*.
Select Relevant Education Materials
Heart Attack Cardiogenic shock Kidney Failure Respiratory Failure

Table 1: Snippet of an EHR note and titles of its linked MedlinePlus documents.

For domain adaptation, we collected Wikipedia articles that are in the Diabetes category. This ensures the Wikipedia articles are from the same domain. The internal Wikipedia links in each article are used as key concepts. There are a total of 130 Wikipedia articles.

The education material collection to evaluate retrieval performance consists of approximately 9400 documents from the "Health Topics", "Drugs and Supplements", and "Medical Encyclopedia" sections of the MedlinePlus website. On average, the documents have 749 tokens, with a standard deviation of 566.

4 Methods

4.1 Domain Adaptation Approaches

We trained Conditional Random Fields (CRF) models to predict the key concepts. As a baseline system, we used leave-one-out cross validation on the EHR notes. The features in the model include lexical, capitalization, prefix, suffix, word shape, and UMLS semantic type. The semantic types are provided by MetaMap, and added as a feature to each token of the MetaMap-recognized terms.

We compared three different methods of domain adaptation to identify the key concepts—instance weighting, instance pruning, and feature augmentation. In accordance with the common terminology, we refer to the larger Wikipedia data as source domain, and the smaller 20 EHR notes the target domain data.

[1] http://www.nlm.nih.gov/medlineplus/

Instance weighting (Jiang and Zhai, 2007) merges the data from both corpora with different weights during training. The weights are usually inversely proportional to the size of the corpus. A model is then trained using this weighted training dataset. In our experiments, we used leave-one-out cross validation on the target domain data. In each fold, the training data is a weighted combination of the Wikipedia data and 19 EHR notes. The test data is the left out EHR note.

Instance pruning (Jiang and Zhai, 2007) removes misleading training instances from the source domain by first applying a model learned from the target domain. For example, if an instance is assigned different labels in the source and target domain corpora, it is removed to prevent the algorithm from learning from this confusing data. We first trained a model on the target domain data, and then predicted the labels on the source domain data. Instances in the source domain that were incorrectly labeled were pruned from the source training set. Finally, a new model was trained using this pruned source domain dataset.

Feature augmentation (Daumé III, 2007) adds additional features to the training instances to identify which corpus they come from. For each original feature in a training example, a new indicator feature is included to indicate the origin domain of the feature, so the learning algorithm can distinguish features important to each domain. A model is then trained on the combined dataset. In our experiments, we applied cross validation on the target domain in a similar fashion to the instance weighting experiments. In each fold, a feature-augmented corpus was built from all the Wikipedia data and 19 EHR notes, and the test data consisted of one EHR note.

4.2 Query Generation

To evaluate the key concepts' effectiveness on education material retrieval, we used the key concepts as queries. The textual MedlinePlus documents are indexed using Galago (Croft et al., 2010), an open source system. In the instance weighting and feature augmentation experiments, the predicted key concepts in the left out EHR note in each fold are combined as queries. In the instance pruning experiments, the predicted key concepts in the EHR notes using the pruned source domain data are used as queries.

Following the same design as reported in Zheng and Yu (2015), we experimented with a two-stage approach, using the same parameters. This approach first issues a query using the key concepts, and then issues a second query using all the concepts recognized by MetaMap. The top 20 results from the first query and the results from the second query are merged to be the final result, removing duplicates between the two result sets.

In all the IR systems, we use Mean Average Precision (MAP) (Manning et al., 2008), a common metric in the IR community, to evaluate the ranked retrieval results. Set-based measures such as precision and recall metrics cannot distinguish the order the results are presented in a ranked retrieval context.

System	Precision	Recall	F1
Baseline	45.77%	26.51%	31.76
Instance Weighting	47.59%	34.41%	38.32
Instance Pruning	40.00%	6.02%	10.23
Feature Augmentation	46.60%	28.86%	34.08

Table 2: Key concept identification results.

5 Results

The results of the baseline CRF model and the models using domain adaptation approaches are shown in Table 2. The baseline system achieved an F1 score of 31.76. Two domain adaptation approaches, instance weighting and feature augmentation, outperformed the baseline system. Both the precision and recall were improved in these two approaches. The best performing approach (instance weighting) shows a 6.56 points (20.7%) improvement in F1 measure over the baseline system.

The Information Retrieval results using these key concepts as queries are shown in the "MAP" column in Table 3. Queries generated from the instance weighting approach outperformed the baseline query results by 0.019 points (20.6%). The other two approaches did not improve over the baseline query.

Results using two-stage approach is shown in the "Two Stage" column of Table 3. The instance weighting approach again outperformed the baseline approach by 0.031 points (27.8%). The other two approaches' performances were similar to the baseline result.

Table 3: Information Retrieval performance. In each system, the queries are generated by combining the recognized key concepts.

Queries	MAP	Two Stage
Baseline	0.0921	0.1114
Instance Weighting	0.1111	0.1424
Instance Pruning	0.0316	0.1002
Feature Augmentation	0.0684	0.1081

6 Discussions

In the domain adaptation experiments, the precision of the three approaches were relatively close to the baseline. However, the recall scores vary greatly. In the instance weighting experiment, the model was able to identify many abbreviations that are rare in the target domain. For example, "EGD" and "DVT" were successfully identified as key concepts despite their occurring only once and three times in the target domain corpus. On the other hand, the instance pruning approach removed over half of the training instances from the source domain data, resulting in a lower performance. The Wikipedia Manual of Style states that only the first occurrence of a term should be linked, and generally a link should only appear once. This resulted in many valid instances being removed because of multiple occurrences. For example, repeated mentions of "glucose" in Wikipedia articles were predicted as key concepts by the target domain model. However, most were removed because only one of them in each article was linked to the glucose article. The reduced training size lowered the recall of this model.

In the IR experiments, the instance weighting approach outperformed the baseline in both the single query and the two stage designs. This can be attributed to the higher recall of this approach in the CRF model. Due to its low recall in key concept identification, instance pruning failed to retrieve many relevant documents. For example, in six of the EHR notes, only one phrase was labeled as key concept, and one of them was incorrect. Despite feature augmentation's improvement in the key concept identification experiments over the baseline, queries generated from this approach did not improve over the baseline query result. The identified key concepts by this method included abbreviations such as "CHF" and general symptoms such as "nausea", which can be associated with a multitude of diseases.

One limitation of the study is that the retrieval gold standard was annotated by one physician. Additional annotators would produce better annotations.

7 Conclusion

It is shown that identifying the key concepts is an effective strategy to generate queries to link EHR notes to education materials. In this study, we explored several domain adaptation approaches to improve key concept identification from EHR notes. The source domain data from Wikipedia enabled the CRF models to learn from more examples. Our experiments have shown that the best setup outperformed a baseline CRF system by 20.7% using data from Wikipedia. Using key concepts recognized from this setup resulted in the best information retrieval performance, a 20.6% improvement over the baseline. Under a two-stage query strategy, retrieval results using these key concepts outperformed the baseline by 27.8%.

Acknowledgments

This work was supported in part by the Award 1I01HX001457 from the United States Department of Veterans Affairs Health Services Research and Development) Program Investigator Initiated Research. The contents do not represent the views of the U.S. Department of Veterans Affairs or the United States Government.

References

Alan R. Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp*, pages 17–21.

Wendy W. Chapman, Prakash M. Nadkarni, Lynette Hirschman, Leonard W. D'Avolio, Guergana K. Savova, and Ozlem Uzuner. 2011. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *Journal of the American Medical Informatics Association*, 18(5):540–543.

Anni R. Coden, Serguei V. Pakhomov, Rie K. Ando, Patrick H. Duffy, and Christopher G. Chute. 2005. Domain-specific language models and lexicons for tagging. *Journal of Biomedical Informatics*, 38(6):422–430.

W. Bruce Croft, Donald Metzler, and Trevor Strohman. 2010. *Search Engines: Information Retrieval in Practice*. Addison-Wesley.

Daniel Dahlmeier and Hwee Tou Ng. 2010. Domain adaptation for semantic role labeling in the biomedical domain. *Bioinformatics*, 26(8):1098–1104.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, June.

Tom Delbanco, Jan Walker, Sigall K. Bell, Jonathan D. Darer, Joann G. Elmore, Nadine Farag, Henry J. Feldman, Roanne Mejilla, Long Ngo, James D. Ralston, Stephen E. Ross, Neha Trivedi, Elisabeth Vodicka, and Suzanne G. Leveille. 2012. Inviting patients to read their doctors' notes: a quasi-experimental study and a look ahead. *Ann. Intern. Med.*, 157(7):461–470, October.

Lorraine Goeuriot, Liadh Kelly, Wei Li, Jo ao Palotti, Pavel Pecina, Guido Zuccon, Allan Hanbury, Gareth Jones, and Henning Müller. 2014. ShARe/CLEF eHealth evaluation lab 2014, task 3: User-centred health Information Retrieval. In *CEUR Workshop Proceedings*, volume 1180, pages 43–61.

Jing Jiang and ChengXiang Zhai. 2007. Instance weighting for domain adaptation in NLP. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 264–271, June.

Liadh Kelly, Lorraine Goeuriot, Hanna Suominen, Tobias Schreck, Gondy Leroy, DanielleL. Mowery, Sumithra Velupillai, Wendy W Chapman, David Martinez, Guido Zuccon, and João Palotti. 2014. Overview of the ShARe/CLEF eHealth evaluation lab 2014. In Evangelos Kanoulas, Mihai Lupu, Paul Clough, Mark Sanderson, Mark Hall, Allan Hanbury, and Elaine Toms, editors, *Information Access Evaluation. Multilinguality, Multimodality, and Interaction*, volume 8685 of *Lecture Notes in Computer Science*, pages 172–191. Springer International Publishing.

Youngjun Kim, Jennifer Garvin, Julia. Heavirland, and Stéphane M. Meystre. 2013. Improving heart failure information extraction by domain adaptation. *Stud Health Technol Inform*, 192:185–189.

Kaihong Liu, Wendy Chapman, Rebecca Hwa, and Rebecca S. Crowley. 2007. Heuristic sample selection to minimize reference standard training set for a part-of-speech tagger. *Journal of the American Medical Informatics Association*, 14(5):641–650.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

National Center for Education Statistics. 2003. National assessment of adult literacy (NAAL).

Balaji Polepalli Ramesh, Rashmi Prasad, Tim Miller, Brian Harrington, and Hong Yu. 2012. Automatic discourse connective detection in biomedical text. *Journal of the American Medical Informatics Association*, 19(5):800–808.

David Wiljer, Sima Bogomilsky, Pamela Catton, Cindy Murray, Janice Stewart, and Mark Minden. 2006. Getting results for hematology patients through access to the electronic health record. *Can Oncol Nurs J*, 16(3):154–164.

Jiaping Zheng and Hong Yu. 2015. Methods for linking EHR notes to education materials. In *Proc. AMIA Summit on Clinical Research Informatics*, pages 209 – 215.