MedLingMap: A growing resource mapping the Bio-Medical NLP field

Marie Meteer, Bensiin Borukhov, Michael Crivaro,
Michael Shafir, Attapol Thamrongrattanarit
{mmeteer, bborukhov, mcrivaro, mshafir, tet}@brandeis.edu
Department of Computer Science
Brandeis University
Waltham, MA 02453, USA

Abstract

The application of natural language processing (NLP) in the biology and medical domain crosses many fields from Healthcare Information to Bioinformatics to NLP itself. In order to make sense of how these fields relate and intersect, we have created "MedLingMap" (www.medlingmap.org) which is a compilation of references with a multi-faceted index. The initial focus has been creating the infrastructure and populating it with references annotated with facets such as topic, resources used (ontologies, tools, corpora), and organizations. Simultaneously we are applying NLP techniques to the text to find clusters, key terms and other relationships. The goal for this paper is to introduce MedLingMap to the community and show how it can be a powerful tool for research and exploration in the field.

1 Introduction

In any field, understanding the scope of the field as well as finding materials relevant to a particular project paradoxically gets more difficult as the field gets larger. This is even more difficult in a field such as Bio-Medical NLP, since it is at the crossroads of multiple disciplines. The drawbacks of keyword search, even using a specific engine such as Google Scholar, are well documented (Stoica et. al 2007) and recent trends in content aggregation and content curation have emerged to attempt to address the problem. Uses of curation range from those in library science to ensure material remain accessible as format and electronic readers change and to make that information more findable (e.g. Peer and Green 2012) to those in marketing to increase revenue by providing more relevant content (Beaulaurier 2012).

However, these approaches still have challenges. Automatic aggregation over a large body of content still provides too many results without additional filtering mechanism. Content curation, which filters content by value and annotates it to ensure higher quality returns, is expensive since annotating large collections of content with the metadata needed to support faceted search and navigation is a huge challenge.

The goal of the work described in this paper is to provide a framework for creating a useful resource tool bounded by the interests of a specific community which can take advantage of automated clustering and keyword extraction techniques and the use of community based annotation through crowd sourcing and social tatting to provide valuable curation. What is an impossible task for a single team because doable when we successfully harness and empower the community.

The MedLingMap site is available at www.medlingmap.org. Currently MedLingMap has over 300 references many of which are annotated according to a set of "contextual" facets (described below). We first provide some use cases for the system and then go into more detail on the content, infrastructure and origins of the system.

We welcome members of the field to join MedLingMap as a curator to help extend the resource. Just email info@medlingmap.org to get an account. Please include your affiliation.

2 Use Cases

MedLingMap was started as a class project in a Brandeis graduate course on NLP in the Medical Domain to provide a means of finding and organizing the publications in the field and as a data source for exploring trends in topics and relation-

Proceedings of the 2012 Workshop on Biomedical Natural Language Processing (BioNLP 2012), pages 140–145,
Montréal, Canada, June 8, 2012.

ships among researchers. While there are many use cases for such a resource, three stand out.

The first is simply the ability to find material that meets very specific criteria. for example, to find papers using "MetaMap" for named entity extraction over clinical data. MedLingMap's growing collection of references and the necessary meta-data to make it useful is well suited to this task.

The second is to support the exploration of an area. If I'm interested in clinical coding, I can select that subject area and am presented with a number of papers. I notice that Phil Resnik is on a number of papers and may want to follow up on his work. I also see many of the papers are tagged with AHIMA, including an entire proceedings that is worth exploring. I select a paper and see the abstract mentions a particular challenge that is also worth following up on. We are in the process of developing a personal "workspace" that will let researchers record searches, annotate findings, and keep a queue of the "next directions" that might be worth following up.

The third use case gets back to one of the original premises of the work, which is that a "map" of a field goes beyond a collection of materials, it also provides context and can be used to see "hot spots" and trends. In order to provide this information and visualization, we have developed a set of tools applying a variety of NLP techniques, such as clustering, topic identification and tf-idf to the content of the papers. This work is described in more detail in (Thamrongrattanarit, et al, 2012).

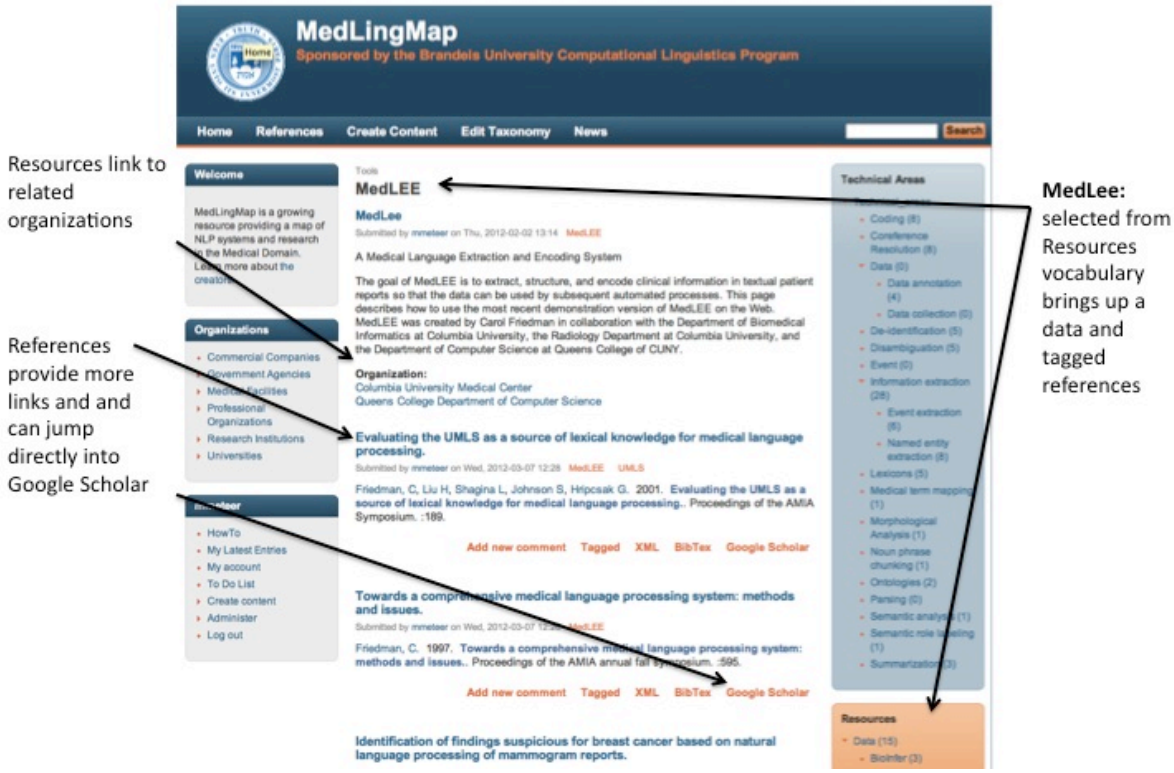


Figure 1: MedLingMap site: www.medlingmap.org

3 Content and Context

The core content in MedLingMap are the references themselves. The underlying representation is based on bibtex and references can be added by either pasting in a single bibtex item or uploading an entire file. A reference can be added through a form interface as well.

We have added BioNLP and related workshops dating back to 2002 as well as many other documents. In addition to the references, there are entries for a variety of organizations and resources

with a short description and links for each. These elements are entered by hand. The assumption is that there are a limited number of them and editorial control is more important than speed of entry.

3.1 Examples of the interface

The MedLingMap interface is shown in Figure 1. All references, resources and organizations are linked through a set of "taxonomies" (described below), which have been developed bottom up based on the material tagged to date. Selecting any item from the taxonomies will select content annotated by that tag. So selecting a "Technical area" from the box on the right brings in all the papers annotated by that topic. A similar box of "resources" allows the user to select all papers that have been annotated as using a particular resource.

For example, in Figure 1 the user has selected MedLEE from the "Resources" taxonomy and is shown the information on MedLee as well as references that have been annotated as discussing MedLEE. In addition to the basic bibliographic information, the user can export the reference in bibtex or xml or jump directly to it Google scholar, which can provide multiple ways of accessing the resource. Alternative views show all of the references by year, author or title.


Biblio

Overview of genia event task in bionlp shared task 2011

View | Edit

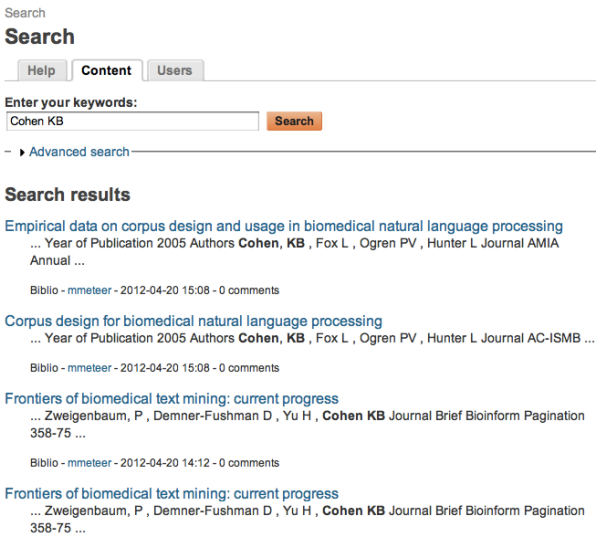
Submitted by mshafir on Tue, 2012-01-31 18:31 BioNLP 2011 Event extraction

Title	Overview of genia event task in bionlp shared task 2011
Publication Type	Journal Article
Year of Publication	2011
Authors	Kim, JD, Wang Y, Takagi T, Yonezawa A
Journal	ACL HLT 2011
Pagination	7
Keywords	BioNLP 2011, Event extraction
Abstract	The Genia event task, a bio-molecular event extraction task, is arranged as one of the main tasks of BioNLP Shared Task 2011. As its sec- ond time to be arranged for community-wide focused efforts, it aimed to measure the advance of the community since 2009, and to evaluate generalization of the technology to full text papers. After a 3-month system de- velopment period, 15 teams submitted their performance results on test cases. The re- sults show the community has made a sig- nificant advancement in terms of both perfor- mance improvement and generalization.
URL	http://www.aclweb.org/anthology/W/W11/W11-1802.pdf

Figure 2: Information on a particular reference

By selecting a reference in MedLingMap, additional information is available, as shown in Figure 2. By selecting any of the key terms from the taxonomy at the top of the "view", the user can go to more papers tagged with that term. By selecting any of the authors, the user is shown other papers by that author. Those with a "curator" account (described below) can select "edit" and make changes or provide additional tags.

In addition, there is a standard search mechanism, as shown in Figure 3. We are in the process of implementing true faceted search, similar to "advanced search" for recipes, where you can select one or more item from each taxonomy to constrain the search.


Search

Search

Help | Content | Users

Enter your keywords:

Cohen KB Search

Advanced search

Search results

Empirical data on corpus design and usage in biomedical natural language processing
... Year of Publication 2005 Authors **Cohen, KB** , Fox L , Ogren PV , Hunter L Journal AMIA Annual ...
Biblio - mmeteer - 2012-04-20 15:08 - 0 comments

Corpus design for biomedical natural language processing
... Year of Publication 2005 Authors **Cohen, KB** , Fox L , Ogren PV , Hunter L Journal AC-ISMB ...
Biblio - mmeteer - 2012-04-20 15:08 - 0 comments

Frontiers of biomedical text mining: current progress
... Zweigenbaum, P , Demner-Fushman D , Yu H , **Cohen KB** Journal Brief Bioinform Pagination 358-75 ...
Biblio - mmeteer - 2012-04-20 14:12 - 0 comments

Frontiers of biomedical text mining: current progress
... Zweigenbaum, P , Demner-Fushman D , Yu H , **Cohen KB** Journal Brief Bioinform Pagination 358-75 ...

Figure 3: Open search

3.2 Faceted indexing

Indexing content along multiple dimensions or "facets" is not new to search (Alan 1995) and significant work has gone into creating effective interfaces for faceted search (Hearst 2006). When searching for research materials, the context the work was done can be a significant contributor to being able to find related materials. "Necessity is the mother of invention" implies that if you want to find similar solutions, look for similar needs.

To try to capture this kind of information, MedLingMap has facets organized into taxonomies:

- Technical area or topic of the work (shown in the screen shot above)
- Resources used:

- Data: Corpora such as Genia, CRAFT, i2B2, BioInfer
- Lexical Resources, which are organized into dictionaries, and ontologies and include UMLS, PubMEd, MedLine, MeSH, and Medical Wordnet
- Tools, such as parsers, taggers, annotation toolkits and more complete systems, such as MedLee, GATE, and MIST
- Shared tasks, such as the BioNLP 2009 and 2011 shared tasks, BioCreative, and i2b2

- Institution the work was done in or is associated with in some way (e.g. funding, providing resources, etc)

As the project continues, these facets will grow and new ones will be added. Additional facets under consideration include the program (e.g. across multiple institutions, generally associated with a single funding source), target data (e.g. medical literature or clinical records).

4 Origins of MedLingMap

As mentioned above, MedLingMap was started as a class project in a graduate course NLP in the Medical Domain and the creation of the taxonomies and population of the material was done as part of the class. However, the underlying architecture itself is based on a system that has been under development for speech recognition for the past two years (www.stcspeechmap.org) by author Marie Meteer as part of the Speech Technology Consortium's effort to improve prior art research in non-patent literature.

The driving principle is that the "art" in any field (the papers, documentation, product descriptions, etc) can only be understood in terms of the context in which they were produced, contexts which show relations between them that is usually not available in the individual documents. For example, much of the early work in speech recognition addressed the challenges of multimodal interfaces well before we had sophisticated mobile devices. Solutions are being reinvented and patents applied for that would not considered novel if the original research were more readily available. Similar issues arise in multidisciplinary fields such as Bio-Medical NLP where different groups come together who do not have the same historical context and may not know about previous research.

5 Infrastructure

MedLingMap and SpeechMap are built on Drupal[1] an open architecture Content Management System (CMS), which underlies many web sites ranging from www.whitehouse.gov to BestBuy.

Using Drupal ensures that MedLingMap can be a living, growing resource. Drupal provides the following functionality:

- A database to store, retrieve, and maintain large documents sets and web pages, providing multiple views into the contents.
- Specific content types for resources, organizations, authors, and references, all linked though a set of taxonomies.
- The capability to load in references in bibtex format either in a group or individually and annotate them with terms from the taxonomies.
- Maintenance facilities, such as suggesting when multiple authors may be the same person and merging them.
- User profiles with different permission levels to accommodate viewers, contributors, social tagging, and private workspaces with the appropriate levels of security.
- The ability to integrate powerful search components, such as SOLR[2], as well as specific modules, such as the Bibliography module which provides automatic links to Google Scholar to retrieve those documents.
- Web-based to allow easy outside access and be more compatible with other systems.
- Extensibility both for more content, more content types, and more functionality. For example while there is a module that produces a warning if a possible duplicate reference, we are still looking for one that would search out potential duplicates and propose merges. If none exists, such a module can be written and easily integrated.

[1] http://drupal.org/
[2] SOLR is an open-source search server based on the Lucene Java search library. http://lucene.apache.org/solr/

6 Value for Stakeholders

The value of MedLingMap varies with the audience. We first talk about the value to the current community and contrast MedLingMap to similar resources already available. We then look at stakeholders outside or entering the community and the value MedLingMap brings to them.

6.1 BioMedical NLP community

For members of the community, a central repository for papers in the field is a "nice to have". There is information that is surfaced by seeing the organization of the information and links to resources in one place, but if you have been attending conferences and workshops regularly, this is not new information. You know the players and already follow the work you are interested in.

In addition, similar information is available elsewhere, though in a more distributed form. ACL has made all of the proceedings to conferences and workshops available[3]. Similarly ACM and IEEE Xplore provide access to all of the papers they control. The significant difference is that in these collections even the advance search is relying on standard bibliographic elements, such as author and title, and keyword search and there is no segmentation of the material by field, which introduces significant ambiguity as the same term can mean different things in different fields. Similarly PubMed and GoPubMed offer documents and advanced search on a huge body of literature, but focused on biology and medicine, not the application of NLP techniques to those fields. MedLingMap is designed to be focused on a smaller community with more like interests.

It is also important to note that MedLingMap is providing links to papers, not the actual papers, which are controlled by the publishers. While many papers are readily available using the links provided or can be found through the Google Scholar link for each reference, if you need a subscription to see the entire paper such as for IEEE, you still need to go through your standard method to get those papers.

LREC's Resource Map is more similar in that it provides more in depth information that the aggregations described above, however the focus is on mapping the resources themselves, not necessarily all of the publications that have taken advantage of those resources, though some of that information may be available by following the links. LREC is also using a crowd sourcing method for growing the resource by asking those who submit papers also submit the information about the resources they used. This is an interesting model in that it assures that those contributing have a stake in the result since they are members of the community by virtue of submitting a paper.

Organizations such as BioNLP.org and SigBioMed are also important resource aggregators for the community. Neither are focused on publications and we hope that MedLingMap willl become one more resource they would point to.

6.2 From the outside

For students or those who come to the field from a neighboring field, the aggregation of the material in MedLingMap can save considerable time and provide overview or "map" of the field. Queries that are ambiguous in Google Scholar are more precise when the domain is limited. <example>

This increase in the ability of newcomers in the field to find what they are looking for actual turns into benefits for those in the field in two ways: First, one's own papers become more findable, increasing citations and potential collaborations. Second, for those who teach, MedLingMap provides a great environment for the students to do targeted research. Letting them loose in a constrained search environment increases the likelihood they will find a rich body of material to learn from and build on without having to always hand select the papers.

7 Growing the resource

The real challenge for a community resource such as MedLingMap is how to grow it to be comprehensive and keep it up to date, specifically how to:

- grow the number of references and resources
- increase high quality annotations that go beyond what can be extracted automatically.
- provide visualizations that bring to light the connections in the material.
- maintain the quality of the data, for example by fining and merging duplicate entries and

[3] http://aclweb.org/anthology-new/

ensuring information about resources and organizations is up to data.

The two choices for growing are automatic techniques and human annotation. We discuss the former in a related paper (Thamrongrattanarit 2012). Here we describe how manual annotation can be feasible.

7.1 Distributed Power

The key to high quality documents and tagging is community involvement. There are two complementary approaches that are key to the MedLingMap project: crowd sourcing and social tagging. Crowd sourcing involves the community in finding relevant resources, particularly those that are fairly obscure and predate the internet. The second is social tagging which lets individuals check on their own materials or materials in areas related to their own work and adding or adjusting the tags to make the content more searchable.

The key to making these tactics work is setting up the right support in the underlying system. Fortunately, the MedLingMap infrastructure allows for easy signup for those volunteering to contribute. These technique have been used successfully in patent prior art search by Article One, Inc.[4] which puts out a call to researchers to find art on a particular patent. If the client selects that art to support their case, the contributor is paid. The patent office itself attempted something similar in the Peer to Patent program[5], which depended on people's desire to improve the quality of patents by letting them contribute art. It was moderately successful, but without the kind of specific reward the Article One provides, they did not get nearly as much material as they would have liked.

MedLingMap, SpeechMap and other efforts of its kind have the same problem: no one has enough time. So how do we address it? How do we create a convincing value proposition? Here are a couple suggestions:

Teaching: MedLingMap is a great teaching tool. Not only can students use it to do research on the material that's in it, we as educators can enlist them to both tag material and go out on the web to find additional material to tag and add. In just one semester we have made considerable progress. If everyone teaching a similar course enlisted their students, the students would gain and the resource would grow.

Research support: With the implementation of the personal workspace described above, the system will provide a unique service not available from other aggregators or content owners.

Funded project: Being able to hire student annotators would accelerate the process. For the SpeechMap project we have a proposal into the US Patent Office for funding. We are open to suggestions about funding sources for MedLingMap.

Conclusion

With MedLingMap's infrastructure in place and enough content to provide an exemplar of how it can grow, the challenge now is engaging the community in what we see as an exiting experiment in harnessing the resources of the internet through crowd sourcing and social tagging to create a living resource that will benefit both the current and future members of the field. MedLingMap also provides a resource for exploring automated ways of annotating and organizing research materials. We also hope that this can be a map itself, to build similar "maps" in other subfields.

References

Allen, RB. 1995. Retrieval from Facet Spaces, Electronic Publishing Chichester, Vol. 8(2 & 3), 247–257.

Beaulaurier, Joe. 2012. Content Curating for Fun and Profit, http://whatcommarketing.com/content-curating-for-fun-and-profit/.

Hearst, Marti. 2006. Design Recommendations for Hierarchical Faceted Search Interfaces. ACM SIGIR Workshop on Faceted Search, August, 2006

Peer, L. Green, A. 2012. Building an Open Data Repository for a Specialized Research Community: Process, Challenges and Lessons, International Journal of Digital Curation, Vol 7, No 1.

E. Stoica, M.A. Hearst, and M. Richardson. 2007. Automating Creation of Hierarchical Faceted Metadata Structures. In Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2007), pages 244–251.

Thamrongrattanarit, A., Shafir, M., Crivaro, M., Borukhov, B. , Meteer, M. What can NLP tell us about Bio NLP? BioNLP Workshop, Montreal, CA, 2012

[4] http://www.articleonepartners.com/

[5] http://peertopatent.org/