Literature-based discovery for Oceanographic climate science

Elias Aamot
Department of Informatics and Computer Science
Norwegian University of Science and Technology
Trondheim, Norway
eliasaa@stud.ntnu.no

Abstract

This paper presents an overview of the field of literature-based discovery, as originally applied in biomedicine. Furthermore it identifies some of the challenges to employing the results of the field in a new domain, namely oceanographic climate science, and elaborates on some of the research that needs to be conducted to overcome these challenges.

1 Introduction

The increase in growth rate of the scientific literature over the past decades has forced researchers to become increasingly specialized in order to keep up with the state of the art. This inevitably leads to the fragmentation of science as researchers from different (sub-)disciplines rarely have time to read each other's papers. Swanson (1986) claimed that this fragmentation of science can lead to *undiscovered public knowledge*: Conclusions that can be made from existing literature, but have never been made because the knowledge fragments have been discovered in separate (sub-)disciplines. Adopting the terminology of Swanson (1991), a literature can be informally defined as a collection of papers with a significant amount of cross-citation related to a single topic. Two literatures are *complementary* if they contain knowledge fragments which can be combined to form new knowledge, and *disjoint* if they have no articles in common, and exhibit little or no cross-citation. The implicit hypothesis is that such *complementary but disjoint* (CBD) literatures are common, giving rise to significant amounts of undiscovered public knowledge. The field of *Literature-based Discovery* (LBD)[1] focuses on the development and application of computational tools to discover undiscovered public knowledge in scientific literature.

[1]Also called *Literature-based knowledge discovery* (LBKD).

Most work in LBD has been conducted in subfields of the biomedical literature, frequently employing knowledge resources specific to that domain. This paper will present an overview of some of the research in LBD, and discuss some of the challenges in reproducing the results made in the LBD field in a different domain, namely oceanographic climate science. The structure of this paper is as follows: Section 2 will give an overview of the LBD field, section 3 will discuss differences between the biomedical domain and that of oceanographic climate science, and section 4 will discuss directions for research that will be conducted in order to adapt LBD methods to the oceanographic climate science domain.

2 Literature-based discovery

Swanson (1986) observed that if a literature L_1 asserted $a \rightarrow b$, and a disjoint literature L_2 asserted $b \rightarrow c$, then the concept denoted by b could function as a bridge between L_1 and L_2, leading to the discovery of the hypothesis $a \rightarrow c$[2]. One example given by Swanson showed that fish oils reduced blood viscosity ($fish\ oil \rightarrow blood\ viscosity$), and that patients of Raynaud's disease tend to exhibit high blood viscosity ($blood\ viscosity \rightarrow Raynaud$). These two facts led to the hypothesis that fish oils can be used in the treatment of Raynaud's disease ($fish\ oil \rightarrow Raynaud$) when combined. This hypothesis was subsequently confirmed experimentally (Digiacomo et al., 1989). Although the inference steps are not logically sound, the procedure is able to produce interesting results. The general approach of bridging dis-

[2]A note on terminology: In the LBD literature, capital letters are normally used for the A, B and C concepts. In this paper, minuscules will be used to represent individual concepts, while capital letters represent sets.

Also, some authors use A to denote the the goal concept, and C for the starting concept. This paper follows the most commonly used terminology, in which a always denotes the starting concept, and c denotes the goal concept.

Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 1–10,
Gothenburg, Sweden, April 26-30 2014.

joint literatures by means of intermediary terms has been dubbed *Swanson linking*, and is also referred to as the *ABC model*.

Swanson and Smalheiser (1997) explain that the discovery of the ABC structure and the fish oil-Raynaud's disease connection happened accidentally. This discovery led Swanson to conduct literature searches aided by existing information retrieval tools to search for more undiscovered public knowledge using the ABC model, resulting in the discovery of eleven connections between migraine and magnesium (Swanson, 1988). As the discovery process was extremely time consuming, requiring the researcher to read hundreds of papers, Swanson later developed a computational tool, Arrowsmith, to streamline the discovery process.

There are two modes of discovery in the ABC model: *Open discovery* and *closed discovery*. In open discovery, the researcher only knows the starting concept a, and is interested in uncovering undiscovered public knowledge related to a. A researcher who looks for consequences of ocean acidification might conduct an open discovery search with $a = ocean\ acidification$. In closed-discovery, the researcher knows both the starting concept a and the goal concept c, and is interested in finding concepts B that prove an explanation of the relationship between the two terms. A researcher who hypothesizes that ocean acidification might cause a reduction in phytoplankton population and tries to discover the causality chain might conduct a closed discovery search with $a = ocean\ acidification, c = phytoplankton\ population$.

This section will present an overview of the state-of-the-art of the LBD field. As this paper discusses the adaptation of LBD to new domains, approaches will be grouped into of three groups according to their dependence on domain specific tools and resources, because reliance on these is likely to hinder cross-domain adaptation[3].

2.1 Group 1: Domain-independent approaches

In the general Swanson linking paradigm, open discovery is conducted by extracting all relations $a \rightarrow b_i$ from the literature of a, written $L(a)$. For every b_i, all relations $b_i \rightarrow c_j$ are then extracted from $L(b_i)$. The set of all $a \rightarrow b_i \rightarrow c_j$ relations, dubbed *discovery candidates* is then are presented to the user as potential discoveries, sorted according to some ranking metric.

In most LBD approaches $L(x)$ is defined as the set of documents returned when searching for x in a literature database. The literature database most commonly used in LBD is Pubmed/Medline[4], maintained by the US National Library of Medicine. The original Arrowsmith system considered only paper titles, as Swanson considered these to hold the most compact knowledge, but it has become the standard approach in LBD to use abstracts and possibly index terms in addition to the titles. The motivation for this is that abstracts and index terms contain more knowledge than only titles.

Somewhat surprisingly, few LBD systems use full paper texts. Schuemie et al. (2004) show that 30-40% of all information contained in a section is new to that section, meaning that significant amounts of knowledge is lost when only looking at abstracts and index terms of a paper. The need for full text data is also pointed out by Cameron et al. (2013). The reason for not using full text seems to be that paper abstracts and index terms are available in xml format through the Pubmed API, while full paper texts require accessing rights and are normally stored as pdf.

In co-occurrence based systems, a relation $x \rightarrow y$ is postulated if x and y exhibit a high degree of co-occurrence in $L(x)$, either in terms of absolute frequency of co-occurrence, or in terms of statistical unlikelihood given the statistical promiscuity of the two concepts. While a few systems use the sentence as the domain for counting co-occurrences, most systems count co-occurrences across entire abstracts.

To present the user with only potential new discoveries, most LBD systems remove from C all terms that are already known to be in a relation with a. In co-occurrence based methods, this is done by removing any (a, c) pairs that exhibit higher degrees if co-occurrence than a predefined threshold (normally 1 co-occurrence) in $L(a)$.

2.1.1 Arrowsmith

The original Arrowsmith system works as follows (Swanson and Smalheiser, 1997): $L(a)$ is fetched

[3] Some of the papers are presented as domain independent, even though they employ domain specific resources, because their main research contributions can be adapted in a domain-independent manner.

[4] `http://www.ncbi.nlm.nih.gov/pubmed/`

by conducting a Medline search to retrieve the titles of papers containing a in the title. The set of potential B concepts is extracted as the list of unique words in $L(a)$, after a stop list of approximately 5000 words has been applied. The B-term set is further pruned by removing all the words that have lesser relative frequency in $L(a)$ than in Medline. The potential B terms are subsequently presented to the user, who can then remove words that are thought to be unsuitable. For each $b_i \in B$, $L(b_i)$ is retrieved and a set C_i is generated, subject to the same stopword and frequency restrictions as before. The terms in the union of the C_i sets are then ranked according to the number of b-terms that connect them to the a-term.

2.1.2 Information retrieval-based methods

Gordon and Lindsay (1996) (Lindsay and Gordon, 1999) developed a system in parallel, which differed from Arrowsmith in several ways: Firstly, while Arrowsmith was word-based, their system used n-grams as the unit of analysis. A stop list was applied by removing all n-grams that contained any stop word occurrence. Secondly, their system used entire Medline records, comprising of keywords, abstracts and titles, whereas Arrowsmith only used paper titles. Thirdly, their system employed information retrieval metrics such as *tf*idf* to find b-terms among the generated candidates, whereas Arrowsmith was based on relative frequencies.

The lexical statistical approach is so generic that it lends itself directly to application in different domains. In a later paper, Gordon et al. (2001) employ this approach to conduct LBD searches directly on the World Wide Web, searching for application areas for genetic algorithms. It should however be noted that the goal of this experiment was not LBD in the sense of uncovering undiscovered public knowledge, instead focusing in discovering something that might be "publicly known" but novel to the user.

2.1.3 Ranking metrics

Wren et al. (2004) pointed out that the structure of concept co-occurrence relationships is such that most concepts are connected to any other concept within few steps. This *small world phenomenon* implies that research focus should be shifted away from retrieving discovery candidates to ranking them, because a significant portion of the concept space will be retrieved even within two co-occurrence relation steps. The paper proposes ranking implicit relationships by comparing the number of observed indirect connections between a and c to the number of expected connections in a random network model, given the relative promiscuity of the intermediary terms.

In another paper, Wren (2004) emphasizes the importance of using a statistically sound method of ranking relationship strengths, such as "chi-square tests, log-likelihood ratios, z-scores or t-scores", because co-occurrence based measures bias towards more general, and thus less interesting relationships. The paper further proposes an extension to the mutual information measure (MIM) as a ranking measure.

2.1.4 Latent semantic indexing

Gordon and Dumais (1998) propose exploiting the ability of certain vector-based semantic models such as Latent semantic indexing (LSI) to discover implicit relationships between terms for LBD. They first train the semantic model on $L(a)$, and let the user choose as b one of the terms most similar to a. A new semantic model is built from $L(b)$, and discovery candidates are ranked according to their similarity to a in the $L(b)$-model. Their experiments showed that the resulting b- and c-term candidate lists closely resemble the lists produced by the information retrieval inspired lexical statistics.

In another experiment they built a semantic model from a random sample of all of Medline, and looked directly for c-terms in the semantic model by considering the terms most similar to a. This "zoomed-out" approach produced different results than the previous Swanson linking inspired approach, which the authors claimed meant that the two methods are complementary and could therefore be used in parallel, but no in-depth evaluation was conducted on the quality of the results.

2.1.5 Evaluation efforts

LBD has a tradition for questionable evaluation effort. The original discoveries in LBD were made manually by Swanson, and most computational systems are evaluated solely according to their ability to replicate one or more of Swanson's discoveries. This is problematic for several reasons: First of all, Swanson's discoveries were never intended as a gold standard, and being able to accomplish a single task that is known in advance does not mean that the results are generalizable.

Secondly, there is no quantitative basis for comparing different approaches or metrics.

Yetisgen-Yildiz and Pratt (2009) conducted the first systematic quantitative evaluation of discovery candidate ranking metrics and relation ranking/generation techniques. They partitioned Medline into two parts, according to a cut-off date. LBD was conducted on the pre-cut-off set, and the post-cut-off set was used as a gold standard to compute precision and recall. In the post-cut-off set, a connection was considered to exist if two terms co-occurred in any document. The ranking metrics that were evaluated were Linking term count (LTC), that is the number of b-terms connecting a and c, Average minimum weight (AMW), that is the average weight of the $a \rightarrow b \rightarrow c$ connections, and Literature cohesiveness (COH), a measure developed by Swanson but not widely adopted. Experiments showed that LTC gave better precision at all levels of recall. The relation generation techniques that were considered were association rules, tf-idf, z-score and MIM. The experiment showed that association rules give the best precision score (8.8%) but the worst recall score (53.76%), while tf-idf gave the best recall (88.0%) but a rather low precision (2.29%).

While the evaluation effort was an important contribution to the LBD field, more quantitative evaluation is required. First of all, all candidate ranking/generation techniques and ranking metrics were tested with only one value of the parameters (for instance the cut-off score for tf-idf, and the cut-off probability for z-score). Comparing the performance of different settings for the parameters would yield a better understanding of each of the metrics, and could lead to results completely different than those reported. Secondly, only a small subset of possible relation generation/ranking techniques and discovery candidate ranking metrics were tested. For example, no relation extraction-based methods (see section 2.3) were included in the evaluation.

The evaluation methodology can be critiqued in several ways. Firstly, building the gold standard from the post-cut-off set is problematic for several reasons: A co-occurrence can exist in the post-cut-off set without necessarily corresponding to a new discovery. Also, as pointed out in Kostoff (2007), it is very difficult to verify that a discovery has not been made before the cut-off date. Another problem is that the post-cut-off set only contains discoveries that have been made in the present, all future discoveries are therefore excluded from the gold standard. Secondly, it is not obvious that quantitative measures reflect the usefulness of the LBD system: When at all is said and done, the usefulness of a LBD system equates to its ability to support user in discovering knowledge.

2.2 Group 2: Concept-based approaches

Several researchers advocate using domain specific concepts taken from an ontology or controlled vocabularies instead of n-gram tokens. Using concepts provides three benefits over n-gram models: Firstly, synonyms and spelling variants are mapped to the same semantic concept. Secondly, using concepts allows for ranking and filtering according to semantic categories. Finally, it becomes easier to constrain the search space by removing spurious or irrelevant n-grams at an early stage, as they don't map to any concept in the domain. On the other hand, concept extraction from raw text is a non-trivial operation.

In LBD concept extraction is conducted in one of two ways: One option is to use NLP tools designed for entity recognition. The most commonly used in the biomedical domain is *MetaMap* (Aronson and Lang, 2010), which extracts concepts from the Unified Medical Language System (UMLS) meta-thesaurus[5]. The other option is to use Medical Subject Headings (MeSH)[6]. MeSH is a controlled vocabulary for indexing biomedical papers, with which all Medline papers have been manually tagged. MeSH keywords can be queried directly from the Medline API. Both MeSH and UMLS terms are organized hierarchically according to semantic categories.

2.2.1 DAD

In their system, DAD (Disease-Adverse reaction-Drug), Weeber et al. (2001) use MetaMap. They showed in an experiment that the number of concepts extracted is significantly lower than the number of n-grams, even after stop lists are applied (8,362 n-grams vs. 5,998 concepts). DAD also allows the user to specify which semantic categories to consider, by for instance only allowing concepts of the type *pharmacological substance* as c concepts, reducing the number of search paths significantly.

[5] http://www.nlm.nih.gov/research/umls/
[6] http://www.nlm.nih.gov/mesh/

Their approach was able to replicate both Swanson's *Raynaud's-fish oil* and *migraine-magnesium* discoveries, but it was discovered that MetaMap maps both *mg* (milligram) and *Mg* (magnesium) to the concept *magnesium*, giving optimistic results for the migraine-magnesium experiment. This is but one example showing that one of the problems with employing NLP tools in an LBD system is that system performance becomes closely tied to the performance of the tools it employs.

2.2.2 LitLinker

Pratt and Yetisgen-Yildiz (2003) developed a system, LitLinker, which originally also used MetaMap, but they later found it too computationally expensive for practical use (Yetisgen-Yildiz and Pratt, 2006). MeSH terms are therefore employed instead.

In a preprocessing step, LitLinker calculates the co-occurrence patterns of every MeSH term across the literatures of every other MeSh term. For every MeSH term, the mean and standard deviation of co-occurrence counts across the literatures is calculated. In the discovery process, a term is considered to be related to another term if their co-occurrence is higher than statistically expected, based on its z-score.

Yetisgen-Yildiz and Pratt identified three classes of uninteresting links and terms that should be pruned automatically by system: (1) too broad terms (giving the examples *medicine*, *disease* and *human*), (2) too closely related terms (giving the example *migraine* and *headache*), and (3) semantically nonsensical connections. The first class is handled by removing any concept if it is strictly more specific in the MeSH ontology hierarchy than any included term. The second class is handled by pruning all links between terms that are closely related (grandparents, parents, siblings and children) in the ontology. The third class is handled by letting the user specify which semantic classes of concepts are allowed to link.

2.2.3 Bitola

Hristovski et al. (2001) originally developed a system called Bitola[7] that discovered *association rules* between MeSH terms. Association rules mining is a common data mining method for discovering relations between variables in a database. Association rules are traditionally used for market basket analysis, in which rules of the type $\{pizza, steak\} \rightarrow \{coca\ cola\}$ are inferred, stating that if somebody buys pizza and steak, he/she is likely to buy coca cola as well. In Bitola's discovery step, basic associations are first mined from the co-occurrence patterns of MeSH terms. Subsequently, indirect associations $a \rightarrow c$ are inferred by combining association rules on the form $a \rightarrow b_i$ and $b_i \rightarrow c$, and ranked according to the sum of strengths of the connecting association rules.

[7] http://ibmi3.mf.uni-lj.si/bitola/

2.3 Group 3: Relation extraction-based approaches

Hristovski et al. (2006) point out two problems with the co-occurrence based LBD systems: Firstly, no explicit explanation of the relation between the a and c terms is given. Secondly, a large number of spurious relations are discovered, as demonstrated by the low precision values witnessed during system evaluation. Both aspects increase the time needed to examine the output of the system by the human user. They suggest that employing natural language processing (NLP) techniques to extract explicit relations from the papers can improve performance on both points.

The biomedical information extraction tool most commonly used in LBD is *SemRep* (Rindflesch and Fiszman, 2003), which uses linguistically motived rules on top of the ouput from MetaMap and the Xerox POS Tagger to extract knowledge in the form of $< subject, predicate, object >$ relation triplets. Although the knowledge expressed in natural language is more complex than what can be represented in simple relation triplets, SemRep is able to provide a better approximation to the knowledge content of scientific papers than do co-occurrence based methods.

While most LBD research employs the same NLP tool, systems differ as to how the extracted relations are represented and how reasoning is conducted in the relation space. Some researchers closely follow the Swanson linking paradigm, and use relation extraction based method instead of or in addition to co-occurrence based methods for candidate generation and ranking. Other researchers take an approach motivated by Wren's observation that a small-world property holds in the network of concept relations in literature. As significant portions of the concept-relation space will have to be explored in a two-step search anyway, it might be better to extract all relations from

the entire literature collection or from a random sample thereof, and rather focus on valid and efficient reasoning within the entire concept-relation space.

Smalheiser (2012) critiques the usage of relation extraction in LBD and claims that while reasoning over explicit relations may lead to so-called *incremental discoveries*, that is, discoveries that lie close to the existing knowledge and therefore are less interesting, they are not able to lead to any *radical discoveries*, that is discoveries that seem unlikely at time of discovery. He also claims that human discoveries, both incremental and radical, tend to be on a higher level, using analogies and abstract similarities rather than explicit relations, and that the benefit from using relation extraction therefore is minimal[8].

2.3.1 Augmented Bitola

In two papers, Hristovski et al. (2006; 2008) experiment with augmenting the Bitola system by using relation extraction tools. In addition to SemRep, they also use another tool, *BioMedLee*, because each of the tools exhibits better performance than the other on certain types of relations.

To guide search through the concept-relation space, they introduce the notion of a *discovery pattern*. A discovery pattern is a set of concept types and relations between them that could imply an interesting relationship in the domain. One discovery pattern, *maybe_treats* can informally be stated as: If a disease leads to a biological change, and a drug leads to the opposite change, then the drug may be able to treat the disease.

The integration between Bitola and the NLP components presented in the system is rather crude; for a given query term, Bitola outputs a set of related terms and the set of papers connecting each related term to the query term. The connecting paper must then be manually input into the NLP components to extract the relation between the query term and any related term. Following a discovery pattern requires extracting relations between several concepts until a chain of the correct relations has been found. The possibility to integrate Bitola and the NLP tools more tightly has been raised as possible future work, but it has been noted a concern that the computational load increases as the NLP component becomes less constrained by the co-occurrence based components.

[8] Smalheiser's critique also extends to many of the widely employed co-occurrence based methods. The argument is that research should focus on developing methods that rank interesting relations highly.

2.3.2 Graph-based reasoning

The extracted relations can be represented as a *Predications Graph* in which each concept is represented by a node and each relation is a labelled, directed edge from the subject concept to the object concept. Representing the concept-relation space as a graph provides two benefits: As a visual tool, a graph can display the knowledge extracted by the system in a way that is easily understood by the user and can be navigated/explored easily. As a mathematical object, one can employ graph theoretic results when developing algorithms for the reasoning process.

In the work of Wilkowski et al. (2011) an initial graph is constructed by querying a pre-compiled database of predications extracted by SemRep from Medline for all relations containing the a concept. The user then incrementally expands the graph by selecting which terms to query relations for from a list of concepts ranked by their degree centrality (i.e. their degree of connectivity in the graph). After graph construction, potential discovery paths are ranked according to summed degree centrality.

Although some work has been conducted in graph-based LBD, seemingly no research has been conducted on LBD in a global, large-scale predications graph derived from all of Medline, or a sample of it.

2.3.3 Predication-based semantic indexing

Cohen et al. (2012a) propose a hyperdimensional computing technique they call *predication-based semantic indexing* (PSI) for efficient representation and reasoning in the concept-relation space. In PSI, concepts and relations are represented as high-dimensional vectors, where the semantic content of a concept's vector is a combination of all the relations it occurs in and all the concepts it is related to, weighted by the frequency of the relation. The system uses SemRep to extract relations from a sample of 8,182,882 Medline records as input to the training process. Inference in this hyperdimensional space can be performed by ordinary vector operations. The paper shows how PSI enables analogical reasoning along the lines of "x is to what as y is to z?" without explicitly traversing the intermediary relation paths between y and z, leading to efficient inference.

The system could originally only infer analogies along a single one of the pathways connecting two concepts x and y. In a later paper Cohen et al. (2012b) expanded the PSI to allow for analogies along multiple pathways, by introducing a vector operation simulating quantum superposition, efficiently reasoning over the entire subgraph connecting x and y. The paper claims that because real world concepts tend to interact through several pathways, literature-based discovery should strive to be able to reason following a similar pattern.

2.4 Approach type hierarchy

From the previous section, it is easy to see the LBD approaches can be divided into a three-level hierarchy according to their dependence on knowledge resources and NLP tools:

Type 1 approaches do not require any knowledge resources: Terms are extracted directly from text, and relations are hypothesized according to co-occurrence patterns. Because all knowledge is extracted directly from text they are completely domain-independent.

Type 2 approaches choose terms from a predefined set of concepts. Co-occurrence patterns are still used to determine relations. The predefined concepts are normally gathered from a domain-specific ontology or vocabulary.

Type 3 approaches use relation extraction tools to extract concepts and relations from text. Because the relations of interest vary widely between domains, domain-specific NLP tools are normally used.

It is evident from the description above that there is a trade-off between reliance on knowledge resources and system performance, as well as a strong correlation between reliance on knowledge resources and domain-dependence. This poses a challenge when adapting LBD approaches to new domains.

3 Domain differences

The current work is a part of a project researching the effects of climate change on the oceanic food web (i.e. who eats who, and how the relative population sizes affect each other) and the biological pump (roughly the ocean's ability to absorb and retain excess atmospheric CO_2). The following section will discuss some of the research issues related to adapting the LBD techniques from the biomedical domain to that of the target domain.

Oceanographic climate science is a cross-disciplinary domain, bringing together researchers from fields such as biology, chemistry, earth science, climate science and oceanography. The cross-disciplinary nature gives rise to an abundance of disjoint literatures, providing strong incentives for LBD. Unfortunately, in a cross-disciplinary domain, scientists from different fields bring their own terminologies and scientific assumptions, creating challenges for LBD work.

While substantial research and engineering effort has gone into the development of NLP tools and computational knowledge sources in the biomedical domain, oceanographic climate science is in this respect under-resourced. To the best of my knowledge, no domain specific NLP tools exist for any sufficiently closely related domain, and although ontologies and controlled vocabularies exist for some of the related disciplines, such as for biology and chemistry, substantial effort is required to identify and combine the desired resources. As a result, it seems unlikely that any of the knowledge intensive (type 2 and 3) LBD methods can be directly applied to oceanographic climate science. Oceanographic climate science also lacks an indexed literature database that covers the entire field, akin to Medline.

Epistemologically there might be a significant difference between the fields: The objects of study (the ocean in oceanographic climate science and the human body in biomedicine) and their processes are quite different, requiring different types of scientific experiments. It therefore seems likely that the structure of the knowledge produced in the different fields might be different. In medicine, experiments can be conducted in a large population of complete systems (human bodies), while in oceanographic experiments must be conducted by sampling subsystems of a single complete system (the ocean). It is therefore not surprising that preliminary observations seem to imply that the results found in oceanographic climate science do not lend themselves to generalization as easily as do those in biomedicine, and that the former have a stronger context dependence (Compare *Eicosapentaenoic acid AFFECTS Vascular constriction* to *Increased labile dissolved organic carbon REDUCES carbon accumulation GIVEN THAT bac-*

teria growth rate is limited). To account for this, text mining tools must be able to extract preconditions as well as relations, or the user must be involved more closely during discovery pattern application to verify that the extracted relations indeed hold true in the same context.

Example discovery patterns for oceanographic climate science have been developed in cooperation with a domain expert, shedding light on some differences between the domains. One research goal of biomedicine is to understand the interactions between domain concepts in order to treat diseases, which is reflected in discovery patterns such as *maybe_treats* (as mentioned in 2.3.1). The discovery patterns developed for oceanographic climate science target the interactions between directional change events (increase or reduce) in quantitative variables, such as *An increase in* CO_2 *causes a decrease in ocean pH*. The types of interactions targeted by these discovery patterns have a more complex structure than the binary relations that define *maybe_treats*. Because most relation extraction tools extract only binary relations, it seems that simply adapting existing relation extraction tools to the domain will not be sufficient.

Ganiz et al. (2006) discusses that LBD lacks a solid theoretic foundation, as most research is applied, rather than theoretical in nature. Although some inquiry has been conducted into the nature of discoveries (Smalheiser, 2012), there is little knowledge about which properties are required to hold in the domain for the LBD methods to be applicable, but the current work assumes that all scientific disciplines are sufficiently similar for LBD methods to be useful.

4 Research directions

The lack of available knowledge resources and NLP tools for the domain makes it hard to directly employ any of the knowledge intensive LBD methods. The development of relation extraction tools for the domain falls outside the scope of the current thesis, and therefore so does the application of type 3 approaches. Instead, the current thesis will focus on bridging the gap between the different terminologies and writing styles caused by different backgrounds in the cross-disciplinary field. To this end, I propose using an unsupervised approach to jointly learn a semantic parser and an ontology from the literature, following the approach of Poon and Domingos (2010).

Poon and Domingos (2009) show that a semantic parser that is able to make non-trivial abstractions from syntactic structure and word usage can be successfully learned in an unsupervised fashion. The system they describe is for instance able to map passive and active form into the same semantic representation and build realistic synonym hierarchies. One challenge that must be addressed is that the current state-of-the-art clusters words based on their argument frames, leading to highly accurate hierarchical clustering of verbs, but lower performance for nouns as these have less diverse argument frames. One research question that will be addressed is how a larger context can be exploited to yield higher performance for nouns.

In an LBD context, the learning process can be seen as bootstrapping a set of concepts for the domain. The resulting system can be considered a hybrid between a type 1 and type 2 approach in terms of the hierarchy defined in 2.4, as it does not use any domain knowledge, but still proposes a set of concepts. A hypothesis that will be evaluated empirically is whether this will provide better results than a pure type 1 system.

The ontology learned by the system can be edited by a domain expert, or combined with ontologies of related fields as they become available, thus providing an elegant interface for integration with domain knowledge in an incremental fashion. The proposed approach will use Markov Logic, a probabilistic extension to first-order logic (FOL), as a knowledge representation language. Background knowledge can therefore easily be incorporated by formulating it as FOL, and the probabilistic aspect enables the system handle contradictions that may occur when combining background knowledge from multiple sources.

The training data set will consist of paper abstracts collected by querying the Mendeley API[9] with a set of keywords that represent the most interesting topics in the domain. The keywords will be developed with the help of a domain expert. As a pre-processing step, the training sentences will be dependency parsed using the Stanford Parser[10]. The proposed LBD system, Houyi[11], will use synonym clusters as concepts, and generate $a \rightarrow b_i$

[9] Mendeley is a web-based reference manager and academic social network that has a large crowd-sourced database of meta-data, such as abstracts, on scientific papers.

[10] `nlp.stanford.edu/software/lex-parser.shtml`

[11] The system is named after a legendary archer in Chinese mythology.

and $b_i \rightarrow c_j$ relation candidates based on td-idf scores. The choice of tf-idf as relation generation/ranking mechanism is motivated by experiments showing that tf-idf gives high recall at the cost of mediocre precision (see section 2.1.5). Because the system is intended to be augmented by relation extraction tools in the future, recall is favoured over precision, as precision is expected to increase in the final version. The discovery candidates are ranked by the number of paths connecting them to a, also motivated by the quantitative experiments described in section 2.1.5.

Houyi will be evaluated quantitatively by comparing performance on a data set divided into training and test data by a cut-off date, following the approach taken by Yetisgen-Yildiz and Pratt (2009). As discussed in section 2.1.5, this is not a perfect evaluation procedure, but it will at least give an indication as to whether unsupervised semantic parsing and ontology building contributes to LBD performance. The baseline system, Sheshou[12], will use the same ranking metric and candidate generation mechanism as Houyi, and uses the NPs extracted by the Stanford Parser as terms.

Development of domain specific ontologies and relation extraction tools is required to apply type 3 LBD methods in the domain. Although outside the scope of the current thesis, it is expected that the resulting semantic parser and ontology can be useful for the development of more sophisticated tools: The semantic parser can function as a preprocessing step for the relation extraction tool by resolving syntactic and synonymic variations. The ontology can be iteratively improved by integrating existing ontologies and human editing, thus providing a point of origin for domain knowledge engineering.

Acknowledgements

This paper is a part of an ongoing master's thesis under the supervision of Pinar Öztürk, with Erwin Marsi as co-supervisor. I am extremely grateful for their feedback and support.

References

Alan R. Aronson and François-Michel M. Lang. 2010. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association : JAMIA*, 17(3):229–236, May.

Delroy Cameron, Olivier Bodenreider, Hima Yalamanchili, Tu Danh, Sreeram Vallabhaneni, Krishnaprasad Thirunarayan, Amit P. Sheth, and Thomas C. Rindflesch. 2013. A graph-based recovery and decomposition of swanson's hypothesis using semantic predications. *J. of Biomedical Informatics*, 46(2):238–251, April.

Trevor Cohen, Dominic Widdows, Roger W. Schvaneveldt, Peter Davies, and Thomas C. Rindflesch. 2012a. Discovering discovery patterns with predication-based Semantic Indexing. *Journal of Biomedical Informatics*, 45(6):1049–1065, December.

Trevor Cohen, Dominic Widdows, Lance Vine, Roger Schvaneveldt, and Thomas C. Rindflesch. 2012b. Many Paths Lead to Discovery: Analogical Retrieval of Cancer Therapies. In *Quantum Interaction*, volume 7620 of *Lecture Notes in Computer Science*, pages 90–101. Springer Berlin Heidelberg.

Ralph A. Digiacomo, Joel M. Kremer, and Dhiraj M. Shah. 1989. Fish-oil dietary supplementation in patients with Raynaud's phenomenon: A double-blind, controlled, prospective study. *The American Journal of Medicine*, 86(2):158–164, January.

Murat C. Ganiz, William M. Pottenger, and Christopher D. Janneck. 2006. Recent Advances in Literature Based Discovery. In *Journal of the American Society for Information Science and Technology*.

Michael D. Gordon and Susan Dumais. 1998. Using latent semantic indexing for literature based discovery. *Journal of the American Society for Information Science and Technology*, 49(8):674–685, June.

Michael D. Gordon and Robert K. Lindsay. 1996. Toward discovery support systems: a replication, re-examination, and extension of Swanson's work on literature-based discovery of a connection between Raynaud's and fish oil. *Journal of the American Society for Information Science and Technology*, 47(2):116–128, February.

Michael Gordon, Robert K. Lindsay, and Weiguo Fan. 2001. Literature Based Discovery on the World Wide Web. In *ACM Transactions on Internet Technology*, pages 261–275, New York, USA. ACM Press.

Dimitar Hristovski, J. Stare, B. Peterlin, and S. Dzeroski. 2001. Supporting discovery in medicine by association rule mining in Medline and UMLS. *Studies in health technology and informatics*, 84(Pt 2):1344–1348.

Dimitar Hristovski, Carol Friedman, Thomas C. Rindflesch, and Borut Peterlin. 2006. Exploiting semantic relations for literature-based discovery. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, pages 349–353.

[12]Mandarin Chinese for "archer", a reference to Arrowsmith.

Dimitar Hristovski, C. Friedman, T. C. Rindflesch, and B. Peterlin. 2008. Literature-Based Knowledge Discovery using Natural Language Processing. In Peter Bruza and Marc Weeber, editors, *Literature-based Discovery*, volume 15 of *Information Science and Knowledge Management*, chapter 9, pages 133–152. Springer, Heidelberg, Germany.

Ronald N. Kostoff. 2007. Validating discovery in literature-based discovery. *Journal of Biomedical Informatics*, 40(4):448–450, August.

Robert K. Lindsay and Michael D. Gordon. 1999. Literature-based discovery by lexical statistics. *Journal of the American Society for Information Science*, pages 574–587.

Hoifung Poon and Pedro Domingos. 2009. Unsupervised semantic parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 1–10, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hoifung Poon and Pedro Domingos. 2010. Unsupervised ontology induction from text. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 296–305, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wanda Pratt and Meliha Yetisgen-Yildiz. 2003. LitLinker: capturing connections across the biomedical literature. In *Proceedings of the 2nd international conference on Knowledge capture*, K-CAP '03, pages 105–112, New York, NY, USA. ACM.

Thomas C. Rindflesch and Marcelo Fiszman. 2003. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *Journal of Biomedical Informatics*, 36(6):462–477, December.

M. J. Schuemie, M. Weeber, B. J. Schijvenaars, E. M. van Mulligen, C. C. van der Eijk, R. Jelier, B. Mons, and J. A. Kors. 2004. Distribution of information in biomedical abstracts and full-text publications. *Bioinformatics*, 20(16):2597–2604, November.

Neil R. Smalheiser. 2012. Literature-based discovery: Beyond the ABCs. *Journal of the American Society for Information Science and Technology*, 63(2):218–224, February.

Don R. Swanson and Neil R. Smalheiser. 1997. An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artificial Intelligence*, 91(2):183–203, April.

Don R. Swanson. 1986. Undiscovered public knowledge. *The Library Quarterly*, 56(2):pp. 103–118.

Don R. Swanson. 1988. Migraine and magnesium: eleven neglected connections. *Perspectives in Biology and Medicine*, 31(4):526–557.

Don R. Swanson. 1991. Complementary structures in disjoint science literatures. In *SIGIR '91: Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 280–289, New York, NY, USA. ACM.

Marc Weeber, Henny Klein, Lolkje T. de Jong van den Berg, and Rein Vos. 2001. Using concepts in literature-based discovery: Simulating Swanson's Raynaud-fish oil and migraine-magnesium discoveries. *Journal of the American Society for Information Science and Technology*, 52(7):548–557.

Bartłomiej Wilkowski, Marcelo Fiszman, Christopher M. Miller, Dimitar Hristovski, Sivaram Arabandi, Graciela Rosemblat, and Thomas C. Rindflesch. 2011. Graph-based methods for discovery browsing with semantic predications. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2011:1514–1523.

Jonathan D. Wren, Raffi Bekeredjian, Jelena A. Stewart, Ralph V. Shohet, and Harold R. Garner. 2004. Knowledge discovery by automated identification and ranking of implicit relationships. *Bioinformatics (Oxford, England)*, 20(3):389–398, February.

Jonathan D. Wren. 2004. Extending the mutual information measure to rank inferred literature relationships. *BMC bioinformatics*, 5, October.

Meliha Yetisgen-Yildiz and Wanda Pratt. 2006. Using statistical and knowledge-based approaches for literature-based discovery. *Journal of Biomedical Informatics*, 39(6):600–611, December.

Meliha Yetisgen-Yildiz and Wanda Pratt. 2009. A new evaluation methodology for literature-based discovery systems. *Journal of biomedical informatics*, 42(4):633–643, August.