Supervised Hypothesis Discovery Using Syllogistic Patterns in the Biomedical Literature

Kazuhiro Seki and Kuniaki Uehara
Kobe University
Japan
seki@cs.kobe-u.ac.jp

Abstract

The ever-growing literature in biomedicine makes it virtually impossible for individuals to grasp all the information relevant to their interests. Since even experts' knowledge is limited, important associations among key biomedical concepts may remain unnoticed in the flood of information. Discovering those hidden associations is called *hypothesis discovery*. This paper reports our approach to this problem taking advantage of a triangular chain of relations extracted from published knowledge. We consider such chains of relations as implicit rules to generate potential hypotheses. The generated hypotheses are then compared with newer knowledge for assessing their validity and, if validated, they are served as positive examples for learning a regression model to rank hypotheses. This framework, called *supervised hypothesis discovery*, is tested on real-world knowledge from the biomedical literature to demonstrate its effectiveness.

1 Introduction

The amount of scientific knowledge is rapidly growing beyond the pace one could digest. For example, Medline,[1] the most comprehensive bibliographic database in life science, currently contains over 19 million references to journal articles and 2,000–4,000 completed references are added each day. Given the substantial volume of the publications, it is virtually impossible for individuals to deal with the information without the aid of intelligent information technologies, such as information extraction [Björne *et al.*, 2010] and text data mining (TDM) [Ananiadou *et al.*, 2006; Kostoff *et al.*, 2009].

TDM aims to discover heretofore unknown knowledge through an automatic analysis on textual data. A pioneering work in TDM, also known as literature-based discovery or hypothesis discovery, was conducted by Swanson in the 1980's. He argued that there were two premises logically connected but the connection had been unnoticed due to overwhelming publications and/or over-specialization. To demonstrate the validity of the idea, he manually analyzed a number of articles and identified logical connections implying a hypothesis that fish oil was effective for clinical treatment of Raynaud's disease [Swanson, 1986]. The hypothesis was later supported by experimental evidence [Digiacomo and adn Dhiraj M. Shah, 1989].

This study is motivated by the series of Swanson's work [Swanson, 1987; 1988; 1990; Swanson *et al.*, 2006] and attempts to advance the research in hypothesis discovery. Specifically, we aim to address two problems that the existing work has generally suffered from. One is the unknown nature of a generated hypothesis. Most approaches only identify two potentially associated concepts, leaving the meaning of the association unknown, which requires experts to interpret the hypothesis. This vagueness has significantly limited the utility of hypothesis discovery. To cope with the problem, we derive hypothesis generation rules from numerous known facts or relations extracted from the scientific literature. Each rule explicitly states the meaning of an association as a predicate and is able to produce an interpretable hypothesis in the form of "N_1 V N_2", where N and V denote a concept (noun phrase) and a predicate (verb phrase), respectively.

The second problem is the large number of generated hypotheses. Typically, most of the hypotheses are spurious and only a small fragment is worth further investigation. Because the latter is far outnumbered by the former and thus is difficult to find, it is crucial to prioritize or rank the hypotheses according to their plausibility. To this end, we first identify true associations among the automatically generated hypotheses and learn their characteristics by adopting a semi-supervised regression model. To build an effective model, we explore several types of features, including the reliability of the hypothesis generation rules, the semantic similarities between concepts, and specificity of concepts. Through a series of experiments on the Medline bibliographic database, we demonstrate the validity of the proposed framework for generation and prioritization of interpretable hypotheses.

2 Related Work

Swanson has argued the potential use of a literature to discover new knowledge that has implicitly existed but been overlooked for years. His discovery framework is based on a syllogism. That is, two premises, "A causes B" and "B causes C," suggest a potential association, "A causes C," where A and C do not have a known, explicit relationship. Such an as-

[1] http://www.ncbi.nlm.nih.gov/entrez

sociation can be seen as a hypothesis testable for verification to produce new knowledge, such as the aforementioned association between Raynaud's disease and fish oil. For this particular example, Swanson manually inspected two groups of articles, one concerning Raynaud's disease and the other concerning fish oil, and identified premises that "Raynaud's disease is characterized by high platelet affregability, high blood viscosity, and vasoconstriction" and that "dietary fish oil reduces blood lipids, platelet affregability, blood viscosity, and vascular reactivity," which together suggest a potential benefit of fish oil for Raynaud's patients. Based on the groundwork, Swanson himself and other researchers developed computer programs to aid hypothesis discovery. The following summarizes some of the representative studies.

Weeber *et al.* [2001] implemented a system, called DAD-system, taking advantage of a natural language processing tool. The key feature of their system is the incorporation of the Unified Medical Language System (UMLS) Metathesaurus[2] for knowledge representation and pruning. While the previous work focused on words or phrases appearing in Medline records for reasoning, DAD-system maps them to a set of concepts defined in the UMLS Metathesaurus using MetaMap [Aronson, 2001]. An advantage of using MetaMap is that it can automatically collapse different wordforms (e.g., inflections) and synonyms to a single Metathesaurus concept. In addition, using *semantic types* (e.g., "Body location or region") under which each concept is categorized, irrelevant concepts can be excluded from further exploration if particular semantic types of interest are given. This filtering step can drastically reduce the number of potential associations, enabling more focused knowledge discovery. Pratt and Yetisgen-Yildiz [2003]'s system, LitLinker, is similar to Weeber's, also using the UMLS Metathesaurus but adopted a technique from association rule mining [Agrawal *et al.*, 1996] to find two associated concepts.

Srinivasan [2004] developed another system, called Manjal, for hypothesis discovery. A primary difference of Manjal from the others is that it solely relies on MeSH[3] terms assigned to Medline records, disregarding all textual information. MeSH is a controlled vocabulary consisting of sets of terms (MeSH terms) used to manually indexing articles in life science. Manjal conducts a Medline search for a given concept and extracts MeSH terms from the retrieved articles. Then, according to predefined mapping, each of the extracted MeSH terms is associated with its corresponding UMLS semantic types. Similar to DAD-system, the subsequent processes can be restricted only to the concepts under particular semantic types of interest, so as to narrow down the potential pathways. In addition, Manjal uses the semantic types for grouping resultant concepts in order to help users browse the system output.

More recently, Liu *et al.* [2011] proposed an approach to hypothesis discovery based on hypergraphs. In the approach, concepts and their direct associations (co-occurrences) are represented by nodes and edges, respectively, and the strength of direct/indirect associations between two concepts were defined using their commute time (taken for a random walk to make a round trip between two nodes) or inner product. They evaluated the approach on synthetic small data, along with shopping basket and clinical note data, and showed that Swanson's hypothesis regarding fish oil can be replicated.

[2]http://www.nlm.nih.gov/research/umls/

[3]http://www.nlm.nih.gov/mesh/

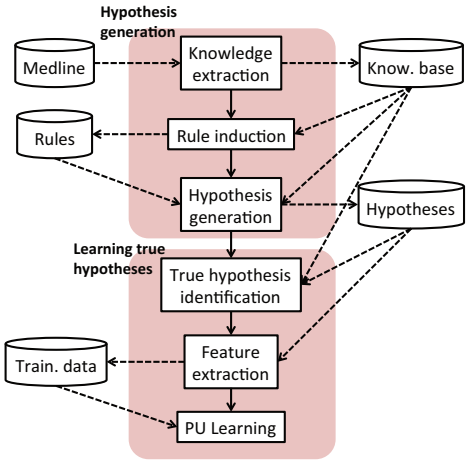

Figure 1: Overview of the supervised hypothesis discovery framework. Solid and dotted lines show the flow of the processes and the flow of the data, respectively.

Despite the prolonged efforts, however, the research in hypothesis discovery is still at an early stage of development, leaving much room to improve. Most of the previous works only suggest two concepts indirectly associated without indicating their nature of the association. Also, their evaluation was typically limited only to a small number of known hypotheses reported in Swanson's work. In contrast, this study proposes a novel approach to generating and ranking hypotheses with explicit meaning and quantitatively evaluates them against a number of known associations automatically extracted from Medline.

3 Proposed Framework

The proposed framework is roughly divided into two parts. One is *hypothesis generation* and the other is *learning true hypotheses*. The former first derives hypothesis generation rules based on known facts or relations represented in predicate argument structure extracted from a corpus of texts and then generates potential hypotheses by applying the acquired rules to known relations. The latter uses true hypotheses as positive examples for learning a regression model and applies it to the generated hypotheses for prioritization according to their plausibility. Figure 1 illustrates the overview of the framework. The following sections describe each component depicted in the figure by roughly following the flow of the processes/data.

3.1 Deriving Hypothesis Generation Rules

Much previous work generates hypotheses simply based on co-occurrences of two terms or concepts. Although such approaches may produce valid hypotheses, they also produce

even more spurious ones, making it more difficult to spot truly important hypotheses. This study takes into account the meaning of the relation between two concepts instead of their simple co-occurrence and only produces more reasonable hypotheses in consideration of the existing knowledge. In this study, each known relation extracted from the existing knowledge is expressed as a predicate-argument structure "N_1 V N_2", where V is a predicate and N_1 and N_2 are subjective and objective arguments, respectively. Based on the same arguments, these relations are merged to identify a *chain of relations* described shortly to derive a hypothesis generation rule.

Knowledge Extraction

To extract known relations from the literature, this study relies on publicly available NLP tools, specifically, a shallow syntactic parser and a named entity (NE) recognizer. Based on the former's output, predicate-argument structure is identified. For example, from a sentence "the guideline is being reexamined currently by the NCRP committee", a relation ⟨the guideline, is being reexamined currently by, the NCRP committee⟩ is extracted. In addition, the following preprocessing is applied in this order to normalize the representation.

- Transform all the terms to their base forms.
- Remove all articles.
- Replace negative adverbs (e.g., barely) with "not"
- Remove all adverbs (except for "not").
- Remove the relation itself if auxiliary verb is uncertain ("may" or "might").
- Remove auxiliary verb.
- Remove present/past/future tense.
- Transform passive voice to active.

For the above example, the extracted relation is finally normalized to ⟨NCRP committee, reexamine, guideline⟩.

An NE recognizer identifies biomedical entities, including proteins and RNA. We retain only relations containing at least one such entity, which would help to generate biomedically meaningful hypotheses. Hereafter, the set of extracted relations is refereed to as the knowledge base, denoted as $\mathbf{K}$.

Rule Induction and Hypothesis Generation

From the knowledge base $\mathbf{K}$, a hypothesis generation rule r is derived as a sequence of three predicates V_1, V_2, V_3. The basic idea is to identify a syllogistic pattern composed of three relations corresponding to two premises and one conclusion in Swanson's syllogism (see Section 2) by merging the same arguments. For example, suppose that two relations were extracted from the literature: "N_1 inhibits N_2" and "N_2 directs N_3". The objective and subjective arguments of the former and latter, respectively, are the same (i.e., N_2) and thus form a chain of two relations by merging them: "N_1-inhibit-N_2-direct-N_3". Here, the previous work in hypothesis discovery may suggest that N_1 and N_3 have *some* implicit association without being able to specify the meaning of the association. In contrast, we take a further step to search for another relation involving N_1 and N_3, such as "N_1 impairs N_3" in the knowledge base $\mathbf{K}$.[4] This time, the subjective and objective arguments are the same as those of the first two relations, respectively. Further merging the arguments produces a triangular chain of relations as shown in Figure 2.

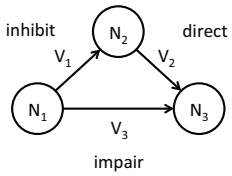

Figure 2: A chain of relations leading to a hypothesis generation rule.

These known relations collectively suggest that a general rule r below may hold:

> **Rule** r**:** If "x inhibits y" and "y directs z", then "x impairs z",

where x, y, and z can be any noun phrases. Note that the rule only indicates a possible association that may be invalid. However, because the possible association follows a more reasonable logic than mere co-occurrence-based approaches, fewer spurious hypotheses would be expected than the previous work.

These rules, denoted as $\mathbf{R} = \{r_1, r_2, \cdots\}$, can be easily identified by first finding two predicate-argument structures that share the same argument as the object and subject (i.e., N_2), and then finding another predicate-argument structure having the other two arguments (N_1 and N_3) as its subject and object. Here, it should be mentioned that rule r also keeps the information on N_1, N_2, and N_3 for reasons described in Section 3.2. Once such rules $\mathbf{R}$ are exhaustively identified in the knowledge base $\mathbf{K}$, they can be applied back to $\mathbf{K}$ to generate hypotheses, denoted as $\mathcal{H} = \{h_1, h_2, \cdots\}$ where h is a generated hypothesis. Let us stress that, in the hypothesis, the exact meaning of the association between two concepts is explicitly stated as a predicate (i.e., "impair" in the above example).

3.2 Learning True Hypotheses

The number of hypotheses $|\mathcal{H}|$ to be generated from the rules $\mathbf{R}$ will be much smaller than co-occurrence-based approaches adopted by the previous work. Still, there will be many hypotheses that hinder manual investigation. Therefore, it is crucial to prioritize or rank the generated hypotheses by considering their plausibility. To this end, we attempt to learn the characteristics of the "true" hypotheses using a supervised learning framework and predict a plausibility or confidence of each generated hypothesis h. Specifically, we take the following three steps: identification of true hypotheses, feature extraction, and PU learning, each described below.

[4] In fact, three relations, "actinomycin D *inhibits* mRNA", "mRNA *directs* protein synthesis", and "actinomycin D *impairs* protein synthesis", were extracted from Medline.

Identification of True Hypotheses

For applying supervised learning, there need to be labeled examples, i.e., true and false hypotheses in this case. Such labeled examples are often manually created in many classification/regression tasks, such as spam filtering. However, it is not realistic to manually judge the validity of the generated hypotheses since it may require domain expertise, extensive reading, and even laboratory experiments. Instead, we take advantage of the biomedical literature more recent than those used for rule induction and hypothesis generation. In other words, if a generated hypothesis is found in the recent literature, the hypothesis can be seen as a validated, true hypothesis.

Specifically, we first split the knowledge base $\mathbf{K}$ into three subsets: $\mathbf{K}_1, \mathbf{K}_2, \mathbf{K}_3$ with $\mathbf{K}_i$ being the ith oldest set of knowledge. Then, $\mathbf{K}_1$ is used for inducing the rules, denoted as $\mathbf{R}_1$, and then for generating hypotheses, denoted as $\mathcal{H}_1$. Among $\mathcal{H}_1$, true hypotheses are identified using more recent knowledge $\mathbf{K}_2$. Note that the remaining, newest knowledge $\mathbf{K}_3$ will be held for evaluation as described later.

A potential problem of this approach is that the hypotheses not found in $\mathbf{K}_2$ cannot be simply regarded as false hypotheses. This is because it is possible that they are actually true hypotheses but not yet appear in the literature. In other words, definite negative examples are difficult to identify. To cope with this issue, we see the non-true, inconclusive hypotheses as "noisy" data containing both true and false hypotheses and adopt a learning approach using positive and unlabeled examples, so called *PU learning*. Specifically, this study adopts an existing approach proposed by Lee and Liu [Lee and Liu, 2003], which considers PU learning as a problem of learning with noise by labeling all unlabeled examples as negative and learns a linear regression model.

Feature Extraction

To apply a supervised learning method, each hypothesis (instance) needs to be represented by a set of features that possibly reflect the characteristics of true/false hypotheses. There are two types of information that can be used to predict the plausibility of hypothesis h generated by a rule r. One is associated with r itself and the other is associated with r and h. In the following, the former is called *rule-dependent* features, and the latter *rule/hypothesis-dependent* features. Note that a rule is represented by a sequence of verbs (V_1, V_2, V_3) but retain the noun phrases (N_1, N_2, N_3) in the syllogistic pattern from which it is derived so as to obtain some of the features.

For the rule-dependent, this work uses the features summarized below.

- The number of syllogistic patterns that resulted in the same rule (V_1, V_2, V_3). If multiple patterns lead to the same rule, it is thought to be more reliable.
- Specificity of verbs. More specific verbs may lead to more specific, useful hypotheses. Following the intuition, two features below are extracted for each verb, V_1, V_2, and V_3, involved in a rule r (see Figure 2).
  - Document frequency (DF) in Medline. The inverse of DF is often used as an indicator of the specificity of a word in information retrieval [Sparck Jones, 1972].
  - The number of synonyms. We assume that broader terms have more synonyms, and *vice versa*. The number is based on an English lexical database, WordNet [Fellbaum, 1998]
- Specificity of nouns. The assumption is similar to the above. For each noun, N_1, N_2, and N_3, involved in a rule r, two features are extracted. When the same r is derived from multiple syllogistic patterns, the generated hypothesis is replicated for each pattern to encode different feature values.
  - DF in Medline.
  - The number of synonyms.

For the rule/hypothesis-dependent features, the following set is utilized.

- The number of rules that produced the same hypothesis h.
- Specificity of nouns involved in the premises "N_4 V_1 N_5" and "N_5 V_2 N_6" that produced h (See the right-hand side of Figure 3). For each noun (N_4, N_5, N_6), its DF in Medline is used as specificity.
- Applicability of rule r to generate hypothesis h. We assume that h is more plausible if r which generated h is more appropriate to the context (two premises) to be applied. We define this applicability of r as "analogical resemblance" between the syllogistic pattern from which r was derived and the one associated with h. The details are described in the next paragraphs.

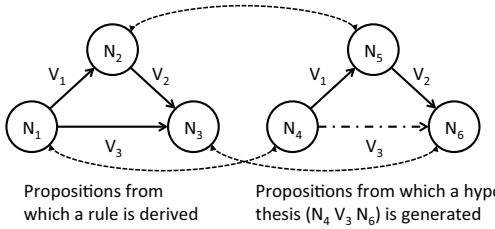

Figure 3: Analogical resemblance between two chains of relations. A dashed line connects two concepts playing the same role in the rule represented by V_1, V_2, and V_3.

Analogical Resemblance

Figure 3 illustrates the idea of analogical resemblance, where the left triangle is the syllogistic pattern from which a rule is derived and the right triangle is the two premises "N_4 V_1 N_5" and "N_5 V_2 N_6" from which a possible hypothesis "N_4 V_3 N_6" is inferred. A dashed line connects two concepts that play the same role in the syllogistic rule represented by a sequence of predicates V_1, V_2, and V_3. If the connected concepts are semantically more similar to each other, the rule is likely to be more applicable to the right triangle with concepts N_4, N_5, and N_6.

There is much work in estimating the semantic similarity of two concepts, such as corpus-based and lexicon-based ap-

proaches [Mihalcea *et al.*, 2006]. This study adopts a corpus-based approach for its wide coverage, specifically, *Normalized Google Distance* (NGD) [Cilibrasi and Vitanyi, 2007]. NGD is an approximation of Normalized Information Distance (NID) and replaces the Kolmogorov complexity in the formulation with the number of Google hits as defined in

$$\mathrm{NGD}(x,y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log N - \min\{\log f(x), \log f(y)\}}, \quad (1)$$

where $f(x)$ is the number of hits by Google search with query x and N is the number of web pages indexed by Google. $\mathrm{NGD}(x,y)$ ranges from 0 to ∞ and $\mathrm{NGD}(x,y)$=0 means that they are identical.

Instead of Google, however, we use Medline, which would better reflect the domain knowledge and also ensures that $f(x)$ for any concept x will exist (non-zero) since the concepts in the propositions are all extracted from Medline in this study. Although the formulation is exactly the same, we call the distance used with Medline *Normalized Medline Distance* (NMD) to avoid unnecessary confusion. It should be mentioned that Lu and Wilbur [2009] also used Medline for computing NGD. We compute an NMD value for each of the three pairs of concepts in Figure 3, i.e., $\mathrm{NGD}(\mathrm{N}_1,\mathrm{N}_4)$, $\mathrm{NGD}(\mathrm{N}_2,\mathrm{N}_5)$, $\mathrm{NGD}(\mathrm{N}_3,\mathrm{N}_6)$, to represent the applicability of rule r to hypothesis h.

4 Evaluation

4.1 Experimental Procedure and Settings

To demonstrate the validity of our proposed framework, we performed evaluative experiments. As the existing knowledge, we used a subset of Medline, specifically, the 2004 Genomics track data set [Hersh *et al.*, 2004]. The data set is a subset of Medline from 1994 to 2003 and is composed of 4,591,008 records. From the data set, known relations were extracted in a predicate-argument structure from the titles and abstracts of the Medline records using the Genia tagger [Tsuruoka and Tsujii, 2005]. After applying the normalizing processes described in Section 3.1, 17,904,002 relations were acquired and formed knowledge base $\mathbf{K}$. Then, $\mathbf{K}$ was split into three subsets $\mathbf{K}_1, \mathbf{K}_2, \mathbf{K}_3$ of around the same size. The oldest knowledge, $\mathbf{K}_1$, was used for rule derivation and hypothesis generation. The number of rules was 12,180 and the number of generated hypotheses was 346,424 including duplicates.

The generated hypotheses were then compared with the knowledge base $\mathbf{K}_2$ and subsequently $\mathbf{K}_3$. If a hypothesis was found in $\mathbf{K}_2$, it was considered as a positive example and was added to the training data. If it was not found in $\mathbf{K}_2$ but found in $\mathbf{K}_3$, it was added to the test data as a positive example. The hypotheses not found in $\mathbf{K}_2$ nor $\mathbf{K}_3$ were unlabeled and were added to either the training or test data at random. This process ensures that training and test data do not have the same instances.

The training data were used for PU learning and the test data were used for evaluating the performance of our proposed framework for hypothesis generation and ranking. The training data contain 226 positive and 169,060 unlabeled examples, and the test data contain 88 positive and 169,059 unlabeled examples. The unlabeled examples in the test data are regarded as negatives in evaluation, although they may not be truly negatives as they may be verified in future. We will come back to this issue in the next section when discussing the evaluation criteria.

In applying the PU learning approach by Lee and Liu [2003], some parameters need to be set. Following their report, the number of iterations (epochs) was set to 500. Similarly, the decay rate and learning rate were experimentally set to 0.05 and 0.00001, respectively, by consulting their report. Although we have tested a few different values on the test set and the results were not very different, better results may be obtained by systematically optimizing the parameters.

4.2 Results and Discussion

Performance of Hypothesis Generation and Ranking

The generated hypotheses in the test data were ranked by the output of the regression model learned from the training data. The performance of the ranking was evaluated by a receiver operating characteristic (ROC) curve and the area under it (AUC). An ROC curve is plotted with x and y axes being false positive and true positive rates, respectively, and is often used for evaluating the performance of classifiers. There are other commonly used evaluation criteria, including accuracy and F-measure. However, accuracy is not suitable for this experiment as the number of positive and negative examples is heavily unbalanced. F-measure is not suitable either because negative examples in the test data may not be actually negatives. An ROC curve is more appropriate in this setting. It basically indicates the degree to which the distributions of positives and negatives are separated, and even if some negatives are actually positives, the influence on the resulting ROC curve is limited if the number of such not-yet-known positives is much smaller than that of negatives. Figure 4 shows the resulting ROC curve, where the AUC was computed as 0.860.

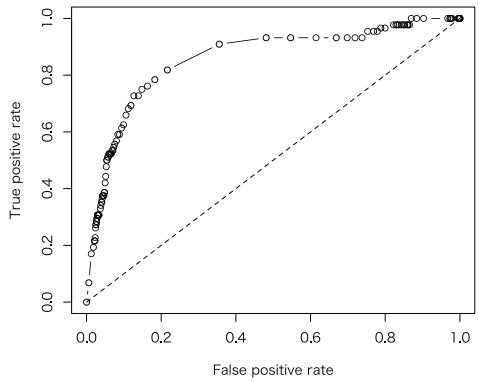

Figure 4: Performance of hypothesis discovery.

As shown by the result, the curve is above the dotted diagonal line, and consequently, AUC is much greater than 0.5 corresponding to random guess, which confirms the overall effectiveness of our proposed framework to hypothesis generation and ranking.

We then looked at which features contributed to the performance based on the regression coefficients (weights) for the

Table 1: Comparison of the features in terms of the regression coefficients in the learned model.

Feature	Regression coefficient	
Rule-dependent	DF of V_3	246.33
	DF of N_2	−176.02
	DF of V_1	150.17
	DF of N_3	−118.27
	DF of N_1	−93.26
	DF of V_2	7.41
	# of patterns for the same rule	4.92
	# of synonyms of V_3	3.70
	# of synonyms of V_1	−2.64
	# of synonyms of N_2	−0.72
	# of synonyms of N_3	−0.17
	# of synonyms of V_2	0.17
	# of synonyms of N_1	−0.05
Rule/hypothesis-dependent	DF of N_4	43.40
	DF of N_6	39.64
	NGD b/w N_1 and N_4	−28.20
	NGD b/w N_3 and N_6	−21.73
	DF of N_5	−14.07
	NGD b/w N_2 and N_5	−9.03
	# of same hypothesis	−0.47

features as summarized in Table 1, where features are sorted in descending order of the absolute weight values within the rule-dependent and rule/hypothesis-dependent groups.

Let us first examine the rule-dependent features. Among them, DF of V_3 was found to have the greatest predictive power with the highest weight of 246.33, followed by DF of N_2 and DF of V_1 and other DF values. The higher effect of V_3's makes sense as it appears as the predicate of the generated hypotheses. Interestingly, DFs for verbs and for nouns were found to have positive and negative weights, respectively. This means that more commonly used general verbs and less commonly used specific nouns tend to form more reliable rules leading to true hypotheses. A possible explanation is that verbs are closed vocabulary and those used to express biological mechanism are often in a regular pattern (e.g., x activates y), whereas nouns in true hypotheses are often specific biomedical entities with lower DFs. Other features including the number of synonyms were found to have little information to predict the true hypotheses.

For the rule/hypothesis-dependent features, DFs of N_4 and N_6 show the higher weights, indicating their positive correlation to true hypotheses. Also, NGD values associated with them were found useful. The negative values of the NGD's weights mean that semantically similar phrases (i.e., smaller distance) lead to true hypotheses, supporting our assumption for analogical resemblance between two syllogistic patterns.

Performance vs. Training Data Size

The training data collected for the evaluation is relatively small, containing only 226 positive examples. Therefore, it is important to investigate the effect of the training data size in learning a regression model to see if more training data would help to increase the performance. For this purpose, we randomly sampled $n\%$ of the training data and used them for learning a regression model. For each n, the same process was repeated for 10 times to compute an average AUC for the particular n. Figure 5 shows the transition of the average AUC with different values of n, where the error bars indicate ± 1 standard deviation.

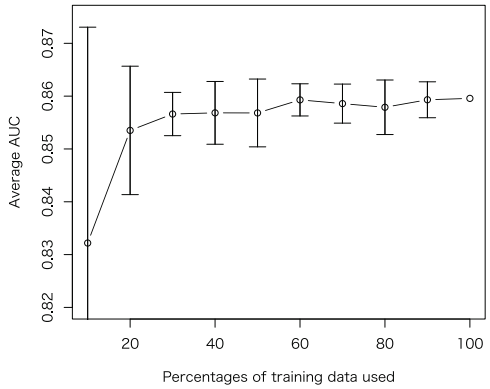

Figure 5: Average AUC for different amount of training data.

The result shows that more training data gradually improve the performance but the difference is subtle. From this experiment, we conclude that more training data would not lead to performance boost with the current model and features. We plan to explore alternative PU learning models [Elkan and Noto, 2008; Xiao *et al.*, 2011, for example] and richer features.

5 Conclusion

This paper focused on a triangular chain of relations, called syllogistic patterns, and proposed a novel approach to hypothesis discovery. The key intuition is that a generalized rule can be induced from such patterns and can then be applied back to the existing knowledge to generate hypotheses. To validate the idea, we implemented the proposed framework and exhaustively identified such hypothesis generation rules in a subset of Medline database and generated hypotheses based on the acquired rules. Among them, true hypotheses were automatically identified based on more recent literature so as to construct training/test data for PU learning. We examined various features associated with specificity, analogical resemblance, and others to represent generated hypotheses. Our evaluation demonstrated that the proposed framework was effective in discovering true hypotheses and that some of those features were characteristic to true hypotheses.

Although the results were promising, the present work has some limitations. The literature used for our experiment was limited in the amount and coverage. Thus, the identified true hypotheses may be actually old knowledge that simply did not appear in our data. Also, basic biomedical knowledge may not appear in Medline and thus not in our knowledge base, either. We plan to address these issues and also explore other PU learning models and features in future work.

Acknowledgments

This work is supported by the Kayamori Foundation grant #K23-XVI-363 and JSPS Kakenhi grant #25330363.

References

[Agrawal *et al.*, 1996] Rakesh Agrawal, Heikki Mannila, Ramakrishnan Srikant, Toivonen Toivonen, and A. Inkeri Verkamo. Fast discovery of association rules. *Advances in knowledge discovery and data mining*, 12:307–328, 1996.

[Ananiadou *et al.*, 2006] Sophia Ananiadou, Douglas B. Kell, and Jun-ichi Tsujii. Text mining and its potential applications in systems biology. *Trends in Biotechnology*, 24(12):571–579, 2006.

[Aronson, 2001] Alan R. Aronson. Effective mapping of biomedical text to the UMLS metathesaurus: The MetaMap program. In *Proceedings of American medical informatics 2001 annual symposium*, pages 17–21, 2001.

[Björne *et al.*, 2010] Jari Björne, Filip Ginter, Sampo Pyysalo, Jun'ichi Tsujii, and Tapio Salakoski. Complex event extraction at PubMed scale. *Bioinformatics*, 26(12):i382–i390, 2010.

[Cilibrasi and Vitanyi, 2007] Rudi L. Cilibrasi and Paul M. B. Vitanyi. The Google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, 19:370–383, 2007.

[Digiacomo and adn Dhiraj M. Shah, 1989] Ralph A. Digiacomo and Joel M. Kremer adn Dhiraj M. Shah. 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, 1989.

[Elkan and Noto, 2008] Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 213–220, 2008.

[Fellbaum, 1998] Christiane D. Fellbaum. *WordNet: an electronic lexical database*. MIT Press, 1998.

[Hersh *et al.*, 2004] William Hersh, Ravi Teja Bhuptiraju, Laura Ross, Aaron M. Cohen, and Dale F. Kraemer. TREC 2004 genomics track overview. In *Proceedings of the 13th Text REtrieval Conference (TREC)*, 2004.

[Kostoff *et al.*, 2009] Ronald N. Kostoff, Joel A. Block, Jeffrey L. Solka, Michael B. Briggs, Robert L. Rushenberg, Jesse A. Stump, Dustin Johnson, Terence J. Lyons, and Jeffrey R. Wyatt. Literature-related discovery. *Annual Review of Information Science and Technology*, 43(1):1–71, 2009.

[Lee and Liu, 2003] Wee Sun Lee and Bing Liu. Learning with positive and unlabeled examples using weighted logistic regression. In *Proceedings of the 20th international conference on machine learning*, pages 448–455, 2003.

[Liu *et al.*, 2011] Haishan Liu, Paea Le Pendu, Ruoming Jin, and Dejing Dou. A hypergraph-based method for discovering semantically associated itemsets. In *Proceedings of the 11th IEEE international conference on data mining*, pages 398–406, 2011.

[Lu and Wilbur, 2009] Zhiyong Lu and W. John Wilbur. Improving accuracy for identifying related PubMed queries by an integrated approach. *Journal of Biomedical Informatics*, 42(5):831–838, 2009.

[Mihalcea *et al.*, 2006] Rada Mihalcea, Courtney Corley, and Carlo Strapparava. Corpus-based and knowledge-based measures of text semantic similarity. In *Proceedings of the 21st national conference on artificial intelligence*, pages 775–780, 2006.

[Pratt and Yetisgen-Yildiz, 2003] Wanda Pratt and Meliha Yetisgen-Yildiz. Litlinker: capturing connections across the biomedical literature. In *Proceedings of the 2nd international conference on Knowledge capture*, pages 105–112, 2003.

[Sparck Jones, 1972] Karen Sparck Jones. Statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–20, 1972.

[Srinivasan, 2004] Padmini Srinivasan. Text mining: generating hypotheses from Medline. *Journal of the American Society for Information Science and Technology*, 55(5):396–413, 2004.

[Swanson *et al.*, 2006] Don R. Swanson, Neil R. Smalheiser, and Vetle I. Torvik. Ranking indirect connections in literature-based discovery: the role of medical subject headings. *Journal of the American society for information science and technology*, 57(11):1427–1439, 2006.

[Swanson, 1986] Don R. Swanson. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, 30(1):7–18, 1986.

[Swanson, 1987] Don R. Swanson. Two medical literatures that are logically but not bibliographically connected. *Journal of the American Society for Information Science*, 38(4):228–233, 1987.

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

[Swanson, 1990] Don R. Swanson. Somatomedin C and arginine: Implicit connections between mutually isolated literatures. *Perspectives in Biology and Medicine*, 33(2):157–179, 1990.

[Tsuruoka and Tsujii, 2005] Yoshimasa Tsuruoka and Jun'ichi Tsujii. Bidirectional inference with the easiest-first strategy for tagging sequence data. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 467–474, 2005.

[Weeber *et al.*, 2001] Marc Weeber, Henry Klein, Lolkje T. W. de Jong-van den Berg, and Rein Vos. 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, 2001.

[Xiao *et al.*, 2011] Yanshan Xiao, Bo Liu, Jie Yin, Longbing Cao, Chengqi Zhang, and Zhifeng Hao. Similarity-based approach for positive and unlabelled learning. In *Proceedings of the 22nd international joint conference on artificial Intelligence*, pages 1577–1582, 2011.