HAL-based Cascaded Model for Variable-Length Semantic Pattern Induction from Psychiatry Web Resources

Liang-Chih Yu and **Chung-Hsien Wu**
Department of Computer Science and Information Engineering
National Cheng Kung University
Tainan, Taiwan, R.O.C.
{lcyu, chwu}@csie.ncku.edu.tw

Fong-Lin Jang
Department of Psychiatry
Chi-Mei Medical Center
Tainan, Taiwan, R.O.C.
jcj0429@seed.net.tw

Abstract

Negative life events play an important role in triggering depressive episodes. Developing psychiatric services that can automatically identify such events is beneficial for mental health care and prevention. Before these services can be provided, some meaningful semantic patterns, such as <lost, parents>, have to be extracted. In this work, we present a text mining framework capable of inducing variable-length semantic patterns from *unannotated* psychiatry web resources. This framework integrates a cognitive motivated model, *Hyperspace Analog to Language (HAL)*, to represent words as well as combinations of words. Then, a cascaded induction process (CIP) bootstraps with a small set of seed patterns and incorporates *relevance feedback* to iteratively induce more relevant patterns. The experimental results show that by combining the HAL model and relevance feedback, the CIP can induce semantic patterns from the unannotated web corpora so as to reduce the reliance on annotated corpora.

1 Introduction

Depressive disorders have become a major threat to mental health. People in their daily life may suffer from some negative or stressful life events, such as death of a family member, arguments with a spouse, loss of a job, and so forth. Such life events play an important role in triggering depressive symptoms, such as depressed mood, suicide attempts, and anxiety. Therefore, it is desired to develop a system capable of identifying negative life events to provide more effective psychiatric services. For example, through the negative life events, the health professionals can know the background information about subjects so as to make more correct decisions and suggestions. Negative life events are often expressed in natural language segments (e.g., sentences). To identify them, the critical step is to transform the segments into machine-interpretable semantic representation. This involves the extraction of key *semantic patterns* from the segments. Consider the following example.

*Two years ago, I **lost** my **parents**.* (Event)

Since that, I have attempted to kill myself several times. (Suicide)

In this example, the semantic pattern <lost, parents> is constituted by two words, which indicates that the subject suffered from a negative life event that triggered the symptom "Suicide". A semantic pattern can be considered as a semantically plausible combination of k words, where k is the length of the pattern. Accordingly, a semantic pattern may have variable length. In Wu et al.'s study (2005), they have presented a methodology to identify depressive symptoms. In this work, we go a further step to devise a text mining framework for variable-length semantic pattern induction from psychiatry web resources.

Traditional approaches to semantic pattern induction can be generally divided into two streams: knowledge-based approaches and corpus-based approaches (Lehnert et al., 1992; Muslea, 1999). Knowledge-based approaches rely on exploiting expert knowledge to design handcrafted semantic patterns. The major limitations of such approaches include the requirement of significant time and effort on designing the handcrafted patterns. Besides, when applying to a new domain, these patterns have to be redesigned. Such limitations form a knowledge acquisition bottleneck. A possible solution to reducing the problem is to use a general-purpose ontology

Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pages 945–952,
Sydney, July 2006.

such as WordNet (Fellbaum, 1998), or a domain-specific ontology constructed using automatic approaches (Yeh et al., 2004). These ontologies contain rich concepts and inter-concept relations such as hypernymy-hyponymy relations. However, an ontology is a static knowledge resource, which may not reflect the dynamic characteristics of language. For this consideration, we instead refer to the web resources, or more restrictedly, the psychiatry web resources as our knowledge resource.

Corpus-based approaches can automatically learn semantic patterns from domain corpora by applying statistical methods. The corpora have to be annotated with domain-specific knowledge (e.g., events). Then, various statistical methods can be applied to induce variable-length semantic patterns from all possible combinations of words in the corpora. However, statistical methods may suffer from data sparseness problem, thus they require large corpora with annotated information to obtain more reliable parameters. For some application domains, such annotated corpora may be unavailable. Therefore, we propose the use of web resources as the corpora. When facing with the web corpora, traditional corpus-based approaches may be infeasible. For example, it is impractical for health professionals to annotate the whole web corpora. Besides, it is also impractical to enumerate all possible combinations of words from the web corpora, and then search for the semantic patterns.

To address the problems, we take the notion of weakly supervised (Stevenson and Greenwood, 2005) or unsupervised learning (Hasegawa, 2004; Grenager et al., 2005) to develop a framework able to bootstrap with a small set of seed patterns, and then induce more relevant patterns form the *unannotated* psychiatry web corpora. By this way, the reliance on annotated corpora can be significantly reduced. The proposed framework is divided into two parts: *Hyperspace Analog to Language (HAL)* model (Burgess et al., 1998; Bai et al., 2005), and a cascaded induction process (CIP). The HAL model, which is a cognitive motivated model, provides an informative infrastructure to make the CIP capable of learning from unannotated corpora. The CIP treats the variable-length induction task as a cascaded process. That is, it first induces the semantic patterns of length two, then length three, and so on. In each stage, the CIP initializes the set of semantic patterns to be induced based on the better results of the previous stage, rather than enumerating all possible combinations of words. This would be helpful to avoid noisy patterns propagating to the next stage, and the search space can also be reduced.

A crucial step for semantic pattern induction is the representation of words as well as combinations of words. The HAL model constructs a high-dimensional context space for the psychiatry web corpora. Each word in the HAL space is represented as a vector of its context words, which means that the sense of a word can be inferred through its contexts. Such notion is derived from the observation of human behavior. That is, when an unknown word occurs, human beings may determine its sense by referring to the words appearing in the contexts. Based on the cognitive behavior, if two words share more common contexts, they are more semantically similar. To further represent a semantic pattern, the HAL model provides a mechanism to combine its constituent words over the HAL space.

Once the HAL space is constructed, the CIP takes as input a seed pattern per run, and in turn induces the semantic patterns of different lengths. For each length, the CIP first creates the initial set based on the results of the previous stage. Then, the induction process is iteratively performed to induce more patterns relevant to the given seed pattern by comparing their context distributions. In addition, we also incorporate expert knowledge to guide the induction process by using *relevance feedback* (Baeza-Yates and Ribeiro-Neto, 1999), the most popular query reformulation strategy in the information retrieval (IR) community. The induction process is terminated until the termination criteria are satisfied.

In the remainder of this paper, Section 2 presents the overall framework for variable-length semantic pattern induction. Section 3 describes the process of constructing the HAL space. Section 4 details the cascaded induction process. Section 5 summarizes the experiment results. Finally, Section 6 draws some conclusions and suggests directions for future work.

2 Framework for Variable-Length Semantic Pattern Induction

The overall framework, as illustrated in Figure 1, is divided into two parts: the HAL model and the cascaded induction process. First of all, the HAL space is constructed for the psychiatry web corpora after word segmentation. Then, each word in HAL space is evaluated by computing its distance to a given seed pattern. A smaller distance represents that the word is more

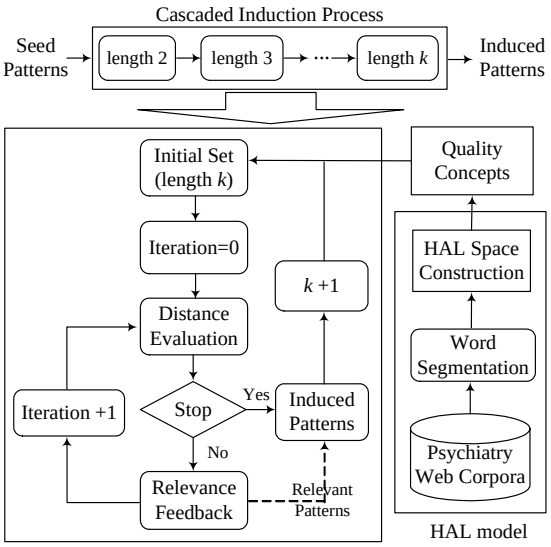

Figure 1. Framework for variable-length semantic pattern induction.

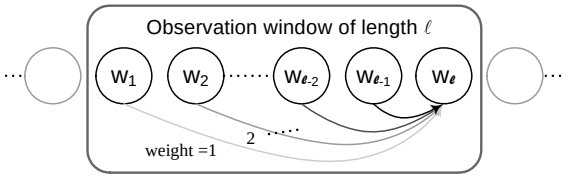

Figure 2. Weighting scheme of the HAL model.

	two	years	ago	I	lost	my	parents
two	0	0	0	0	0	0	0
years	5	0	0	0	0	0	0
ago	4	5	0	0	0	0	0
I	3	4	5	0	0	0	0
lost	2	3	4	5	0	0	0
my	1	2	3	4	5	0	0
parents	0	1	2	3	4	5	0

Table 1. Example of HAL Space (window size=5)

semantically related to the seed pattern. According to the distance measure, the CIP generates *quality concepts*, i.e., a set of semantically related words to the seed pattern. The quality concepts and the better semantic patterns induced in the previous stage are combined to generate the initial set for each length. For example, in the beginning stage, i.e., length two, the initial set is the all possible combinations of two quality concepts. In the later stages, each initial set is generated by adding a quality concept to each of the better semantic patterns. After the initial set for a particular length is created, each semantic pattern and the seed pattern are represented in the HAL space for further computing their distance. The more similar the context distributions between two patterns, the closer they are. Once all the semantic patterns are evaluated, the *relevance feedback* is applied to provide a set of relevant patterns judged by the health professionals. According to the relevant information, the seed pattern can be refined to be more similar to the relevant set. The refined seed pattern will be taken as the reference basis in the next iteration. The induction process for each stage is performed iteratively until no more patterns are judged as relevant or a maximum number of iteration is reached. The relevant set produced at the last iteration is considered as the result of the semantic patterns.

3 HAL Space Construction

The HAL model represents each word in the vocabulary using a vector representation. Each dimension of the vector is a weight representing the strength of association between the target word and its context word. The weights are computed by applying an observation window of length l over the corpus. All words within the window are considered as co-occurring with each other. Thus, for any two words of distance d within the window, the weight between them is computed as $l-d+1$. Figure 2 shows an example. The HAL space views the corpus as a sequence of words. Thus, after moving the window by one word increment over the whole corpus, the HAL space is constructed. The resultant HAL space is an $N \times N$ matrix, where N is the vocabulary size. In addition, each word in the HAL space is called a concept. Table 1 presents the HAL space for the example text "*Two years ago, I lost my parents*."

3.1 Representation of a Single Concept

For each concept in Table 1, the corresponding row vector represents its left context information, i.e., the weights of the words preceding it. Similarly, the corresponding column vector represents its right context information. Accordingly, each concept can be represented by a pair of vectors. That is,

$$c_i = (v_{c_i}^{left}, v_{c_i}^{right}) = \left(\left\langle w_{c_i t_1}^{left}, w_{c_i t_2}^{left}, \ldots, w_{c_i t_N}^{left} \right\rangle, \left\langle w_{c_i t_1}^{right}, w_{c_i t_2}^{right}, \ldots, w_{c_i t_N}^{right} \right\rangle \right), \quad (1)$$

where $v_{c_i}^{left}$ and $v_{c_i}^{right}$ represent the vectors of the left context information and right context information of a concept c_i, respectively, $w_{c_i t_j}$ denotes

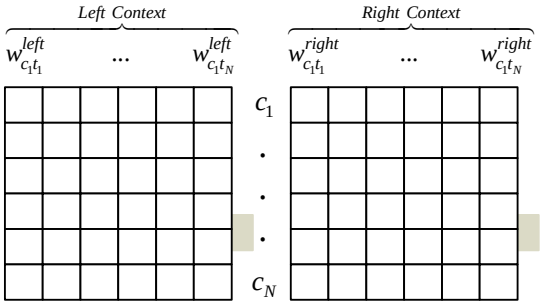

Figure 3. Conceptual representation of the HAL space.

the weight of the *j*-th dimension (t_j) of a vector, and *N* is the dimensionality of a vector, i.e., vocabulary size. The conceptual representation is depicted in Figure 3.

The weighting scheme of the HAL model is frequency-based. For some extremely infrequent words, we consider them as noises and remove them from the vocabulary. On the other hand, a high frequent word tends to get a higher weight, but this does not mean the word is informative, because it may also appear in many other vectors. Thus, to measure the informativeness of a word, the number of the vectors the word appears in should be taken into account. In principle, the more vectors the word appears in, the less information it carries to discriminate the vectors. Here we use a weighting scheme analogous to TF-IDF (Baeza-Yates and Ribeiro-Neto, 1999) to re-weight the dimensions of each vector, as described in Equation (2).

$$w_{c_it_j} = w_{c_it_j} * \log \frac{N_{vector}}{vf(t_j)}, \tag{2}$$

where N_{vector} denotes the total number of vectors, and $vf(t_j)$ denotes the number of vectors with t_j as the dimension. After each dimension is re-weighted, the HAL space is transformed into a probabilistic framework. Accordingly, each weight can be redefined as

$$w_{c_it_j} \equiv P(t_j \mid c_i) = \frac{w_{c_it_j}}{\sum_j w_{c_it_j}}, \tag{3}$$

where $P(t_j \mid c_i)$ is the probability that t_j appears in the vector of c_i.

3.2 Concept Combination

A semantic pattern is constituted by a set of concepts, thus it can be represented through concept combination over the HAL space. This forms a new concept in the HAL space. Let $sp = (c_1,...,c_S)$ be a semantic pattern with *S* constituent concepts, i.e., length *S*. The concept combination is defined as

$$\oplus c_s \equiv ((...(c_1 \oplus c_2) \oplus c_3) \oplus ... \oplus c_S), \tag{4}$$

where $\oplus$ denotes the symbol representing the combination operator over the HAL space, $\oplus c_s$ denotes a new concept generated by the concept combination. The new concept is the representation of a semantic pattern, also a vector representation. That is,

$$\begin{aligned} \oplus c_s &= (v^{left}_{\oplus c_s}, v^{right}_{\oplus c_s}) \\ &= \left(\left\langle w^{left}_{(\oplus c_s)t_1}, \ldots, w^{left}_{(\oplus c_s)t_N} \right\rangle, \left\langle w^{right}_{(\oplus c_s)t_1}, \ldots, w^{right}_{(\oplus c_s)t_N} \right\rangle\right), \end{aligned} \tag{5}$$

The combination operator, $\oplus$, is implemented by the product of the weights of the constituent concepts, described as follows.

$$\begin{aligned} w_{(\oplus c_s)t_j} &= \prod_{s=1}^{S} w_{c_st_j} \\ &= \prod_{s=1}^{S} P(t_j \mid c_s), \end{aligned} \tag{6}$$

where $w_{(\oplus c_s)t_j}$ denotes the weight of the *j*-th dimension of the new concept $\oplus c_s$.

4 Cascaded Induction Process

Given a seed pattern, the CIP is to induce a set of relevant semantic patterns with variable lengths (from 2 to *k*). Let $sp_{seed} = (c_1,...,c_R)$ be a seed pattern of length *R*, and $sp = (c_1,...,c_S)$ be a semantic pattern of length *S*. The formal description of the CIP is presented as

$$\begin{aligned} & sp_{seed} \mid\!- \{sp\} \\ & \equiv (c_1,...,c_R) \mid\!- \{(c_1,...,c_S)\} \quad \text{iff } \forall\, Dist(\oplus c_r, \oplus c_s) \le \lambda, \end{aligned} \tag{7}$$

where $\mid\!-$ denotes the symbol representing the cascaded induction, $\oplus c_r$ and $\oplus c_s$ are the two new concepts representing sp_{seed} and sp, respectively, and $Dist(\bullet, \bullet)$ represents the distance between two semantic patterns. The main steps in the CIP include the *initial set generation*, *distance measure*, and *relevance feedback*.

4.1 Initial Set Generation

The initial set for a particular length contains a set of semantic patterns to be induced, i.e., the search space. Reducing the search space would be helpful for speeding up the induction process,

especially for inducing those patterns with a larger length. For this purpose, we consider that the words and the semantic patterns similar to a given seed pattern are the better candidates for creating the initial sets. Therefore, we generate quality concepts, a set of semantically related words to a seed pattern, as the basis to create the initial set for each length. Thus, each seed pattern will be associated with a set of quality concepts. In addition, the better semantic patterns induced in the previous stage are also considered. The goodness of words and semantic patterns is measured by their distance to a seed pattern. Here, a word is considered as a quality concept if its distance is smaller than the average distance of the vocabulary. Similarly, only the semantic patterns with a distance smaller than the average distance of all semantic patterns in the previous stage are preserved to the next stage. By the way, the semantically unrelated patterns, possibly noisy patterns, will not be propagated to the next stage, and the search space can also be reduced. The principles of creating the initial sets of semantic patterns are summarized as follows.

- In the beginning stage, the aim is to create the initial set for the semantic patterns with length two. Thus, the initial set is the all possible combinations of two quality concepts.
- In the latter stages, each initial set is created by adding a quality concept to each of the better semantic patterns induced in the previous stage.

4.2 Distance Measure

The distance measure is to measure the distance between the seed patterns and semantic patterns to be induced. Let $sp = (c_1, ..., c_S)$ be a semantic pattern and $sp_{seed} = (c_1, ..., c_R)$ be a given seed pattern, their distance is defined as

$$Dist(sp, sp_{seed}) = Dist(\oplus c_s, \oplus c_r), \quad (8)$$

where $Dist(\oplus c_s, \oplus c_r)$ denotes the distance between two semantic patterns in the HAL space. As mentioned earlier, after concept combination, a semantic pattern becomes a new concept in the HAL space, which means the semantic pattern can be represented by its left and right contexts. Thus, the distance between two semantic patterns can be computed through their context distance. Equation (8) thereby can be written as

$$Dist(sp, sp_{seed}) = Dist(v^{left}_{\oplus c_s}, v^{left}_{\oplus c_r}) + Dist(v^{Right}_{\oplus c_s}, v^{Right}_{\oplus c_r}). \quad (9)$$

Because the weights of the vectors are represented using a probabilistic framework, each vector of a concept can be considered as a probabilistic distribution of the context words. Accordingly, we use the *Kullback-Liebler (KL) distance* (Manning and Schütze, 1999) to compute the distance between two probabilistic distributions, as shown in the following.

$$D(v_{\oplus c_s} \| v_{\oplus c_r}) = \sum_{j=1}^{N} P(t_j | \oplus c_s) \log \frac{P(t_j | \oplus c_s)}{P(t_j | \oplus c_r)}, \quad (10)$$

where $D(\bullet \| \bullet)$ denotes the KL distance between two probabilistic distributions. When Equation (10) is ill-conditioned, i.e., zero denominator, the denominator will be set to a small value (10^{-6}). For the consideration of a symmetric distance, we use the divergence measure, shown as follows.

$$Div(v_{\oplus c_s}, v_{\oplus c_r}) = D(v_{\oplus c_s} \| v_{\oplus c_r}) + D(v_{\oplus c_r} \| v_{\oplus c_s}). \quad (11)$$

By this way, the distance between two probabilistic distributions can be computed by their KL divergence. Thus, Equation (9) becomes

$$Dist(v_{\oplus c_s}, v_{\oplus c_r}) = Div(v^{left}_{\oplus c_s}, v^{left}_{\oplus c_r}) + Div(v^{Right}_{\oplus c_s}, v^{Right}_{\oplus c_r}). \quad (12)$$

After each semantic pattern is evaluated, a ranked list is produced for relevance judgment.

4.3 Relevance Feedback

In the induction process, some non-relevant semantic patterns may have smaller distance to a seed pattern, which may decrease the precision of the final results. To overcome the problem, one possible solution is to incorporate expert knowledge to guide the induction process. For this purpose, we use the technique of relevance feedback. In the IR community, the relevance feedback is to enhance the original query from the users by indicating which retrieved documents are relevant. For our task, the relevance feedback is applied after each semantic pattern is evaluated. Then, the health professionals judge which semantic patterns are relevant to the seed pattern. In practice, only the top *n* semantic patterns are presented for relevance judgment. Finally, the semantic patterns judged as relevant are considered to form the relevant set, and the others form the non-relevant set. According to the relevant and non-relevant information, the seed pattern can be refined to be more similar to the relevant set, such that the induction process can induce more relevant patterns and move away from noisy patterns in the future iterations.

The refinement of the seed pattern is to adjust its context distributions (left and right). Such adjustment is based on re-weighting the dimensions of the context vectors of the seed pattern. The dimensions more frequently regarded as relevant patterns are more significant for identifying relevant patterns. Hence, such dimensions of the seed pattern should be emphasized. The significance of a dimension is measured as follows.

$$Sig(t_k) = \frac{\sum_{\oplus c_i \in R} w_{(\oplus c_i)t_k}}{\sum_{\oplus c_j \in \overline{R}} w_{(\oplus c_j)t_k}}, \quad (13)$$

where $Sig(t_k)$ denotes the significance of the dimension t_k, $\oplus c_i$ and $\oplus c_j$ denote the semantic patterns of the relevant set and non-relevant set, respectively, and $w_{(\oplus c_i)t_k}$ and $w_{(\oplus c_j)t_k}$ denote the weights of t_k of $\oplus c_i$ and $\oplus c_j$, respectively. The higher the ratio, the more significant the dimension is. In order to smooth $Sig(t_k)$ to the range from zero to one, the following formula is used:

$$Sig(t_k) = \frac{1}{1 + \left(\sum_{\oplus c_i \in R} w_{(\oplus c_i)t_k} \Big/ \sum_{\oplus c_j \in \overline{R}} w_{(\oplus c_j)t_k} \right)^{-1}}. \quad (14)$$

The corresponding dimension of the seed pattern $sp_{seed} = \oplus c_r$ is then re-weighted by

$$w_{(\oplus c_r)t_k} = w_{(\oplus c_r)t_k} + Sig(t_k). \quad (15)$$

Once the context vectors of the seed pattern are re-weighted, they are also transformed into a probabilistic form using Equation (3). The refined seed pattern will be taken as the reference basis in the next iteration. The relevance feedback is performed iteratively until no more semantic patterns are judged as relevant or a maximum number of iteration is reached. At the same time, the induction process for a particular length is also stopped. The whole CIP process is stopped until the seed patterns are exhausted

5 Experimental Results

To evaluate the performance of the CIP, we built a prototype system and provided a set of seed patterns. The seed patterns were collected by referring to the well-defined instruments for assessing negative life events (Brostedt and Pedersen, 2003; Pagano et al., 2004). A total of 20 seed patterns were selected by the health professionals. Then, the CIP randomly selects one seed pattern per run without replacement from the seed set, and iteratively induces relevant patterns from the psychiatry web corpora. The psychiatry web corpora used here include some professional mental health web sites, such as PsychPark (http://www.psychpark.org) (Bai, 2001) and John Tung Foundation (http://www.jtf.org.tw).

In the following sections, we describe some experiments to in turn examine the effect of using relevance feedback or not, and the coverage on real data using the semantic patterns induced by different approaches. Because the semantic patterns with a length larger than 4 are very rare to express a negative life event, we limit the length *k* to the range of 2 to 4.

5.1 Evaluation on Relevance Feedback

The relevance feedback employed in this study provides the relevant and non-relevant information for the CIP so that it can refine the seed pattern to induce more relevant patterns. The relevance judgment is carried out by three experienced psychiatric physicians. For practical consideration, only the top 30 semantic patterns are presented to the physicians. During relevance judgment, a majority vote mechanism is used to handle the disagreements among the physicians. That is, a semantic pattern is considered as relevant if any two or more physicians judged it as relevant. Finally, the semantic patterns with majority votes are obtained to form the relevant set.

To evaluate the effectiveness of the relevance feedback, we construct three variants of the CIP, *RF(5)*, *RF(10)*, and *RF(20)*, implemented by applying the relevance feedback for 5, 10, and 20 iterations, respectively. These three CIP variants are then compared to the one without using the relevance feedback, denoted as *RF(—)*. We use the evaluation metric, *precision at 30 (prec@30)*, over all seed patterns to examine if the relevance feedback can help the CIP induce more relevant patterns. For a particular seed pattern, prec@n is computed as the number of relevant semantic patterns ranked in the top *n* of the ranked list, divided by *n*. Table 2 presents the results for *k*=2.

The results reveal that the relevance feedback can help the CIP induce more relevant semantic patterns. Another observation indicates that applying the relevance feedback for more iterations

	RF(—)	*RF(5)*	*RF(10)*	*RF(20)*
prec@30	0.203	0.263	0.318	0.387

Table 2. Effect of applying relevance feedback for different number of iterations or not.

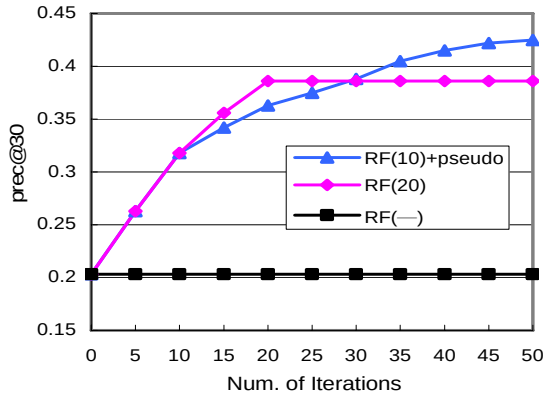

Figure 4. Effect of using the combination of relevance feedback and pseudo-relevance feedback.

can further improve the precision. However, it is usually impractical for experts to involve in the guiding process for too many iterations. Consequently, we further consider *pseudo-relevance feedback* to automate the guiding process. The pseudo-relevance feedback carries out the relevance judgment based on the assumption that the top ranked semantic patterns are more likely to be the relevant ones. Thus, this approach usually relies on setting a threshold or selecting only the top *n* semantic patterns to form the relevant set. However, determining the threshold is not trivial, and the threshold may be different with different seed patterns. Therefore, we apply the pseudo-relevance feedback only after certain expert-guided iterations, rather than applying it throughout the induction process. The notion is that we can get a more reliable threshold value by observing the behavior of the relevant semantic patterns in the ranked list for a few iterations.

To further examine the effectiveness of the combined approach, we additionally construct a CIP variant, *RF(10)+pseudo*, by applying the pseudo-relevance feedback after 10 expert-guided iterations. The threshold is determined by the physicians during their judgments in the 10-th iteration. The results are presented in Figure 4.

The precision of *RF(10)+pseudo* is inferior to that of *RF(20)* before the 25-th iteration. Meanwhile, after the 30-th iteration, *RF(10)+pseudo* achieves higher precision than the other methods. This indicates that the pseudo-relevance feedback can also contribute to semantic pattern induction in the stage without expert intervention.

5.2 Coverage on Real Data

The final results of the semantic patterns are the relevant sets of the last iteration produced by *RF(10)+pseudo*, denoted as SP_{CIP}. Parts of them are shown in Table 3.

Seed Pattern	< boyfriend, argue >
Induced Patterns	<girlfriend, break up>; <friend, fight>

Table 3. Parts of induced semantic patterns.

We compare SP_{CIP} to those created by a corpus-based approach. The corpus-based approach relies on an annotated domain corpus and a learning mechanism to induce the semantic patterns. Thus, we collected 300 consultation records from the PsychPark as the domain corpus, and each sentence in the corpus is annotated with a negative life event or not by the three physicians. After the annotation process, the sentences with negative life events are together to form the training set. Then, we adopt *Mutual Information* (Manning and Schütze, 1999) to learn variable-length semantic patterns. The mutual information between *k* words is defined as

$$MI(w_1,...,w_k) = P(w_1,...,w_k)\log\frac{P(w_1,...,w_k)}{\prod_{i=1}^{k} P(w_i)} \quad (16)$$

where $P(w_1,...w_k)$ is the probability of the *k* words co-occurring in a sentence in the training set, and $P(w_i)$ is the probability of a single word occurring in the training set. Higher mutual information indicates that the *k* words are more likely to form a semantic pattern of length *k*. Here the length *k* also ranges from 2 to 4. For each *k*, we compute the mutual information for all possible combinations of words in the training set, and those with their mutual information above a threshold are selected to be the final results of the semantic patterns, denoted as SP_{MI}. In order to obtain reliable mutual information values, only words with at least the minimum number of occurrences (>5) are considered.

To examine the coverage of SP_{CIP} and SP_{MI} on real data, 15 human subjects are involved in creating a test set. The subjects provide their experienced negative life events in the form of natural language sentences. A total of 69 sentences are collected to be the test set, of which 39 sentences contain a semantic pattern of length two, 21 sentences contain a semantic pattern of length three, and 9 sentences contain a semantic pattern of length four. The evaluation metric used is *out-of-pattern (OOP)* rate, a ratio of unseen patterns occurring in the test set. Thus, the OOP can be defined as the number of test sentences containing the semantic patterns not occurring in the training set, divided by the total number of sentences in the test set. Table 4 presents the results.

	k=2	k=3	k=4
SP_{CIP}	0.36 (14/39)	0.48 (10/21)	0.44 (4/9)
SP_{MI}	0.51 (20/39)	0.62 (13/21)	0.67 (6/9)

Table 4. OOP rate of the CIP and a corpus-based approach.

The results show that the OOP of SP_{MI} is higher than that of SP_{CIP}. The main reason is the lack of a large enough domain corpus with annotated life events. In this circumstance, many semantic patterns, especially for those with a larger length, could not be learned, because the number of their occurrences would be very rare in the training set. With no doubt, one could collect a large amount of domain corpus to reduce the OOP rate. However, increasing the amount of domain corpus also increases the amount of annotation and computation complexity. Our approach, instead, exploits the quality concepts to reduce the search space, also applies the relevance feedback to guide the induction process, thus it can achieve better results with time-limited constraints.

6 Conclusion

This study has presented an HAL-based cascaded model for variable-length semantic pattern induction. The HAL model provides an informative infrastructure for the CIP to induce semantic patterns from the unannotated psychiatry web corpora. Using the quality concepts and preserving the better results from the previous stage, the search space can be reduced to speed up the induction process. In addition, combining the relevance feedback and pseudo-relevance feedback, the induction process can be guided to induce more relevant semantic patterns. The experimental results demonstrated that our approach can not only reduce the reliance on annotated corpora but also obtain acceptable results with time-limited constraints. Future work will be devoted to investigating the detection of negative life events using the induced patterns so as to make the psychiatric services more effective.

References

R. Baeza-Yates and B. Ribeiro-Neto. 1999. *Modern Information Retrieval*. Addison-Wesley, Reading, MA.

Y. M. Bai, C. C. Lin, J. Y. Chen, and W. C. Liu. 2001. Virtual Psychiatric Clinics. *American Journal of Psychiatry*, 158(7):1160-1161.

J. Bai, D. Song, P. Bruza, J. Y. Nie, and G. Cao. 2005. Query Expansion Using Term Relationships in Language Models for Information Retrieval. In *Proc. of the 14th ACM International Conference on Information and Knowledge Management*, pages 688-695.

E. M. Brostedt and N. L. Pedersen. 2003. Stressful Life Events and Affective Illness. *Acta Psychiatrica Scandinavica*, 107:208-215.

C. Burgess, K. Livesay, and K. Lund. 1998. Explorations in Context Space: Words, Sentences, Discourse. *Discourse Processes*. 25(2&3):211-257.

C. Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

T. Grenager, D. Klein, and C. D. Manning. 2005. Unsupervised Learning of Field Segmentation Models for Information Extraction. In *Proc. of the 43th Annual Meeting of the ACL*, pages 371-378.

T. Hasegawa, S. Sekine, R. Grishman. 2004. Discovering Relations among Named Entities from Large Corpora. In *Proc. of the 42th Annual Meeting of the ACL*, pages 415-422.

W.Lehnert, C. Cardie, D. Fisher, J. McCarthy, E. Riloff, and S. Soderland. 1992. University of Massachusetts: Description of the CIRCUS System used for MUC-4. In *Proc. of the Fourth Message Understanding Conference*, pages 282-288.

C. Manning and H. Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press. Cambridge, MA.

I. Muslea. 1999. Extraction Patterns for Information Extraction Tasks: A Survey. In *Proc. of the AAAI-99 Workshop on Machine Learning for Information Extraction*, pages 1-6.

M. E. Pagano, A. E. Skodol, R. L. Stout, M. T. Shea, S. Yen, C. M. Grilo, C.A. Sanislow, D. S. Bender, T. H. McGlashan, M. C. Zanarini, and J. G. Gunderson. 2004. Stressful Life Events as Predictors of Functioning: Findings from the Collaborative Longitudinal Personality Disorders Study. *Acta Psychiatrica Scandinavica*, 110:421-429.

M. Stevenson and M. A. Greenwood. 2005. A Semantic Approach to IE Pattern Induction. In *Proc. of the 43th Annual Meeting of the ACL*, pages 379-386.

C. H. Wu, L. C. Yu, and F. L. Jang. 2005. Using Semantic Dependencies to Mine Depressive Symptoms from Consultation Records. *IEEE Intelligent System*, 20(6):50-58.

J. F. Yeh, C. H. Wu, M. J. Chen, and L. C. Yu. 2004. Automated Alignment and Extraction of Bilingual Domain Ontology for Cross-Language Domain-Specific Applications. In *Proc. of the 20th COLING*, pages 1140-1146.