Intention Classification for Retrieval of Health Questions

Rey-Long Liu*

ARTICLE INFO

Article history:
Received 29 October 2016
Revised 17 February 2015
Accepted 26 February 2017

Keywords:
Health Questions,
Intention Recognition,
Text Classification,
Learning-Based Classifiers,
Location-based Feature Weighting,
Area-based Feature Weighting

ABSTRACT

Healthcare professionals have edited many health questions (HQs) and their answers for healthcare consumers on the Internet. The HQs provide both readable and reliable health information, and hence retrieval of those HQs that are relevant to a given question is essential for health education and promotion through the Internet. However, retrieval of relevant HQs needs to be based on the recognition of the *intention* of each HQ, which is difficult to be done by predefining syntactic and semantic rules. We thus model the intention recognition problem as a text classification problem, and develop two techniques to improve a learning-based text classifier for the problem. The two techniques improve the classifier by *location-based* and *area-based* feature weightings, respectively. Experimental results show that, the two techniques can work together to significantly improve a Support Vector Machine classifier in both the recognition of HQ intentions and the retrieval of relevant HQs.

1. Introduction

The Internet has been a main channel from which healthcare consumers get health information for health management and promotion. Health information is often about specific disorders (Shuyler & Knight, 2003), and it needs to be *reliable*, because it is often used to make health-related decisions (Liszka et al., 2006). Inaccurate information may be perceived as reliable information (Abbas et al., 2010). Moreover, health information also needs to be both *relevant* to users' needs (Eysenbach & Köhler, 2002; Zeng et al., 2004) and *readable* for healthcare consumers (Thomson & Hoffman-Goetz, 2007; Eysenbach & Köhler, 2002; Zeng et al., 2004).

Therefore, In order to provide healthcare consumers with the reliable, relevant, and readable health information, healthcare professionals have edited and posted many health questions (HQs) and their answers on many websites. The HQs are thus a valuable resource for health education and promotion through the Internet. A question (*q*) from a healthcare consumer and retrieval of those HQs that are relevant to *q* is thus a key to promote the utility of valuable health information embedded in the HQ databases.

Table 1 shows several examples of Chinese CHQs. An HQ is often about a disorder plus specific-

* Professor, Department of Medical Informatics, Tzu Chi University Hualien, Taiwan (rlliutcu@mail.tcu.edu.tw)
International Journal of Knowledge Content Development & Technology, 7(1): 101-120, 2017.
http://dx.doi.org/10.5865/IJKCT.2017.7.1.101

intentions asking for certain kinds of information about the disorder. As the HQs are often asked by healthcare consumers, their intentions are often related to the management of specific disorders. Typical intentions include (1) before a disorder is diagnosed: *prevention* of the disorder; (2) when a disorder is being diagnosed: *diagnosis*, *risk factors* and *symptoms* of the disorder; (3) after a disorder is confirmed: *treatment*, *medicine*, and *homecare* for the disorder; and (4) after a disorder is treated: *prognosis*, *mortality*, and *recurrence* of the disorder.

Table 1. Several Chinese HQs from health information websites: An HQ is often about a disorder plus a specific *intention* whose recognition is essential but difficult to be determined by predefined syntactic and semantic rules.

Health Question (in Chinese)	Health Question (in English)	Intention
(1) 溢奶(milk regurgitation)及(and)吐奶(milk vomiting)	milk regurgitation and milk vomiting	*General description*
(2) 關於 (about)黑眼圈(black eye)	About the black eye	*General description*
(3) 我有(I have)牙周病(periodontal disease),醫師(doctor)說(said)我應做(I need)進一步的治療(further treatment), 嚴重(serious)的地方(cases)還需要(need)開刀(surgery), 有必要嗎(is it necessary)?	I get periodontal disease. The doctor said that I need further treatment, and in some serious cases I even need surgery. Is it necessary?	*Medicine*
(4) 我(I)很愛乾淨(love cleanness), 天天(every day)洗澡(bathe), 家裡(home)也打掃(sweep)得很清潔(clean), 家中(home)也不(not)潮濕(wet), 為何(why)我(I)還會患(get)疥瘡(scabies)?	I love cleanness, and bathe every day. My home is swept, kept clean, and not wet. Why do I get scabies?	*Risk factor*
(5) 關於(about)關節炎(arthritis)有沒有什麼(is there any)營養(nutrition)治療(treatment)良方(good way)?	Is there any good nutrition way to treat arthritis?	*Homecare*
(6) 痔瘡(hemorrhoids)開刀(surgery)後(after), 還會(can)復發(recur)嗎	Can hemorrhoids recur after surgery?	*Recurrence*
(7) 如何(how)防範(prevent)及(and)治療(treat)放射線治療(radiation therapy)引起(incur)之口內(mouth)併發症(complications)?	How to prevent and treat the mouth complications incurred by radiation therapy?	*Prevention*; *Medicine*

Recognition of the intention of an HQ is the fundamental basis of retrieving relevant answers to the HQ(Zhang et al., 2016; Liu & Lin, 2012; Cartright et al., 2011). It has also been essential for the retrieval of professional clinical information as well (Lin & Demner-Fushman, 2006); however, the intention recognition is challenging due to several reasons:

(1) An HQ is not always well-formed and can even be composed of single terms (e.g., disorder names, see the 1st HQ in Table 1), a single phrase (e.g., a prepositional phrase, see the 2nd HQ in Table 1) or multiple statements (e.g., statements describing current status of a patient, see the 3rd and the 4th HQs in Table 1).
(2) Many HQs are quite short and hence certain intention-indicative terms are essential for the intention recognition; however the intention-indicative terms often confuse the classifier as

well. For example, "關於" (about) tends to indicate that the user is asking for general description about a disorder (e.g., the 2ndHQ in Table 1), but it can simply happen to appear in another HQ whose intention is totally different (e.g., the 5thHQ in Table 1). As another example, consider "開刀" (surgery). It tends to indicate that the user is asking for medicine information (e.g., the 3rdHQ in Table 1); however it can be simply used to describe a patient case in an HQ asking for the information about the recurrence of a disorder (e.g., the 6thHQ in Table 1).

(3) An HQ can be composed of any terms that are *not* intention-indicative (e.g., disorder names only, see the 1st HQ in Table 1). An HQ can also have multiple intentions (e.g., the 7th HQ in Table 1).

Because of these issues, recognition of the intentions of the HQs is difficult to accomplish by predefined rules (e.g., syntactic and semantic rules). It can be approached by text classifiers that are built by machine learning methodologies; however the classifiers need to be further improved to tackle the above challenges.

In this paper, we investigate the intentions of thousands of Chinese HQs on the Internet. We then model the intention recognition problem as a text classification problem, and develop two techniques to improve a learning-based text classifier for the problem. The two techniques improve the classifier by *location-based* and *area-based* feature weightings, respectively.

The *location-based* feature weighting technique is motivated by an observation: a word in an HQ may be more intention-indicative if it appears at the beginning or the end of the HQ. For example, two words "復發" (recur) at the end of the 6th example in Table 1 provide a strong evidence indicating the HQ should be classified into the *recurrence* category, even though the HQ also mentions two words "開刀" (surgery), which are indicative for another category (the *medicine* category). By properly setting the weights (feature values) of the words at certain locations in an HQ, the classifier maybe improved. Note that the location-based word feature weighting cannot be perfect either. For example, two words "關於" (about) appear at the beginning of the 2nd and the 5th HQs in Table 1; however the two HQs have totally different intentions. The *area-based* feature weighting technique is therefore developed to further improve the classifier.

The *area-based* feature weighting technique is motivated by an observation: an HQ tends to have an intention *c* if it has several words that (1) are *exclusively* indicative for *c* (indicative for *c* but no other categories), and (2) appear in an *area* of the HQ (i.e., the words appear in a nearby area of the HQ). For example, consider the 5th HQ in Table 1. The words "營養(nutrition) 治療(treatment) 良方(good way)"(good way for nutrition treatment) together strongly indicate that the HQ should be classified into the *homecare* category, but the words "治療"(treatment) also suggest classifying the HQ into the *medicine* category (since the words often appear in those HQs that ask for medical care information for disorders). Therefore, by amplifying the weights of "營養"(nutrition) and reducing the weights of "治療"(treatment), the classifier can be improved.

The two feature weighting techniques are thus based on observations on real-world HQs. These observations can thus be justified by investigating the practical contributions of the two techniques. Experimental results show that, the two techniques can seamlessly work together to significantly

improve a representative learning-based question classifier in two experiments: (1) recognition of HQ intentions and (2) retrieval of relevant HQs. The feature weights generated by the two techniques can thus be expected to be helpful in improving question classifiers. The contributions are of practical significance to health promotion and education through the Internet.

2. Related Work

Our main contributions include (1) investigation of the intentions of Chinese HQs on the Internet, and (2) development of the two techniques that improve learning-based intention classifiers by location-based feature weighting and area-based feature weighting. To our knowledge, no previous studies focused on these contributions.

Intentions of HQs are often related to health information inquiry in several scenarios of disorder management: (1) *before* a disorder is diagnosed (e.g., prevention of the disorder), (2) *when* a disorder is being diagnosed (e.g., diagnosis of the disorder), (3) *after* a disorder is confirmed (e.g., treatment of the disorder), and (4) *after* a disorder is treated (e.g., prognosis of the disorder). HQ intentions are actually a type of semantics that is particularly essential in healthcare, and hence previous studies have developed several taxonomies of the intentions for various purposes of retrieving health information (Zhang et al., 2016; Palotti et al., 2014; Liu & Lin, 2012; Cartright et al., 2011; Lin & Demner-Fushman, 2006). A typical usage of the HQ intentions is to retrieve relevant HQs whose answers can provide both readable and reliable information for healthcare consumers (Liu & Lin, 2012). Previous question retrieval techniques often retrieved questions by considering question similarities in question types (e.g., interrogative types "what," "how," and "where," Wu et al., 2005) and syntactic or semantic structures of the questions (Casellas et al., 2007; Wang et al., 2009; Wu et al., 2006). Similarities in the question types and the question structures cannot indicate the similarity in HQ intentions, which is a key to retrieve relevant HQs.

One possible way to recognize the intention of an HQ is to parse and analyze the HQ; however, it is quite difficult to predefine a complete set of rules to analyze the intention of the HQ. Because of this, a previous study employed string pattern matching to recognize the HQ intentions (Liu & Lin, 2012); however, it is still both difficult and costly to predefine a complete set of string patterns for each intention category.

We decided to approach the intention recognition problem by machine learning-based text classifiers, without needing to predefine any rules or patterns. Support Vector Machine (SVM) is a representative-machine learning technique to build question classifiers. It was routinely employed in question classification (e.g., Raghavi et al., 2015; Rekha et al., 2014; Moschitti et al., 2011; Lin et al., 2006; Krishnan et al., 2005) and shown to be one of the best question classifiers (Mishra et al., 2013; Pan et al., 2008; Huang et al., 2008; Zhang & Lee, 2003). These SVM question classifiers were often trained with those features that were derived from syntactic and semantic structures of the questions. It is, however, quite difficult to develop a parser to derive the syntactic and semantic structures for Chinese HQs, due to three reasons: (1) parsing Chinese questions is still a challenging task (Lee et al., 2008), (2) HQs are not always well-formed for parsing (as noted above, they

may even consist of a single term, single phrases, or multiple statements), and (3) there is no Chinese dictionary that can cover all possible health-related terms entered by healthcare consumers.[1)]

It is, therefore, of both technical and practical significance to develop novel techniques to further improve the classifiers for the recognition of Chinese HQ intentions. We thus aim at developing two novel feature weighting techniques (*location-based* feature weighting and *area-based* feature weighting) to improve the question classifier. Both techniques considered the *locality* of *intention-indicative* words in an HQ, rather than other kinds of information such as part-of-speech tags of words (Zhang et al., 2016) and language-specific ontology of medical terms (Palotti et al., 2014). To our knowledge, no previous studies investigated the development and contribution of the two techniques in classifying HQ intentions.

The location-based feature weighting technique derives the weights of features (words) based on the location of the words in the HQs. To our knowledge, no previous techniques have employed word locations to improve question classification. On the other hand, the area-based feature weighting technique derives the weights of features (words) based on how intention-indicative words appear in nearby areas. It is, therefore, related to those techniques that employed term proximity to improve text classification. The previous techniques often considered multiple consecutive terms (i.e., *n*-gram, Peng & Schuurmans, 2003), nearby terms in a fixed order (Cohen & Singer, 1996), co-occurring terms in whatever order and location (Cohen & Singer, 1996), and term-category correlation types of nearby terms (Liu, 2010); however, they aimed at the classification of *general texts*, rather than HQs, which are often quite short. To classify HQ intentions, the previous techniques require substantial revisions due to three reasons: (1) terms in a short HQ are always very close to each other, making term proximity alone unable to indicate HQ intentions, (2) an HQ is often quite short with multiple terms related to different intention categories, and (3) training of the classifiers needs to be based on a limited amount of training data (since HQs are often much shorter than general texts). No previous question classifiers were trained with term proximity information. Our area-based word features are specifically developed for the classification of HQ intentions. Area-based words features for an intention category *c* are those that appear in an area and are *exclusively* indicative for *c* (indicative for *c* but not other categories). We will show that the location-based and area-based word features can significantly improve the classification of HQ intentions.

3. Intentions of Chinese HQs on the Internet: A Survey

We have surveyed thousands of Chinese HQs on the Internet. There were 2171 HQs comprehensively collected from 98 sources, which fell into 7 types: governments, health organizations, hospitals, clinics, health news, health information providers, and blogs of healthcare professionals. We found that the HQs are often motivated by the management of a specific disorder (health condition). Since users that are concerned with a disorder will have different information needs in different stages

1) We employed a parser (http://parser.iis.sinica.edu.tw/) to parse 38 Chinese HQs and found that 34.2% of them cannot have a single and correct parse tree.

of the disorder, we identify four typical stages of a disorder, including: (1) *before* the disorder is diagnosed, (2) *when* the disorder is being diagnosed, (3) *after* the disorder is confirmed, and (4) *after* the disorder is treated. Based on these stages, we develop a hierarchy of HQ intentions. As shown in Figure 1, there are 12 intention categories, including one most-general category (*general description*), four general categories (*prevention*, *diagnosis*, *treatment*, and *prognosis*), and seven specific categories (*risk factors*, *symptoms and signs*, *lab test*, *homecare*, *medicine*, *mortality*, and *recurrence*).

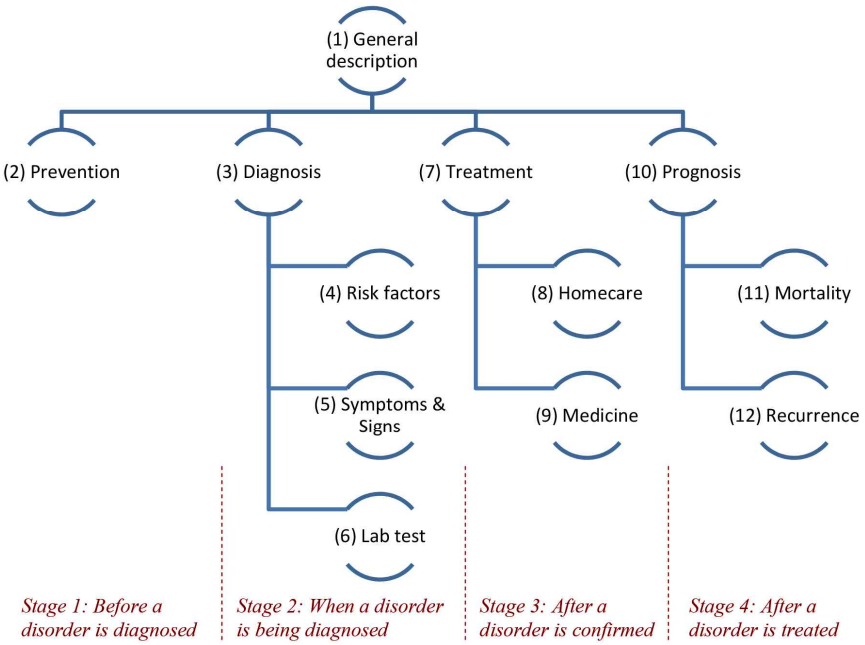

Fig. 1. A hierarchy of HQ intentions: There are 12 intention categories, and specific intentions are motivated in four typical stages of disorders: *before* a disorder is diagnosed (Prevention), *when* a disorder is being diagnosed (Diagnosis), *after* a disorder is confirmed (Treatment), and *after* a disorder is treated (Prognosis).

Based on the hierarchical intention taxonomy, we analyzed the intention of each HQ. The intention of each HQ was identified by two independent people. To ensure the reliability of the intention identification, the two people employed a *focused group* approach in determining the intention of each HQ. They independently assigned an intention category to each HQ solely based on the content of the HQ, and if they assigned different categories to an HQ, a group discussion was conducted so that the two people knew each other's judgment and jointly came up with the final intention category. If the two people still had conflicting judgments for an HQ, another person was invited to join the focused group discussion to finalize the intention category of the HQ.

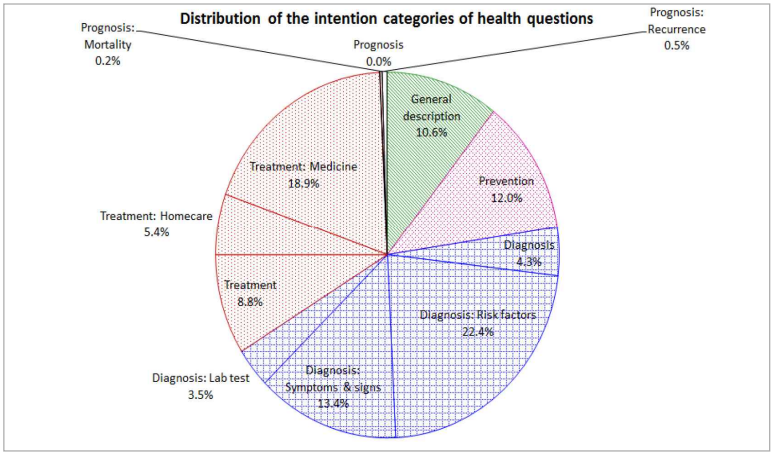

Fig. 2. Distribution of the intention categories of Chinese HQs. The most popular intentions are about the risk factors, symptoms and signs, and medical care (medicine) of disorders.

Figure 2 shows the distributions of the 12 intentions. The results indicated that the HQs were compiled for specific intentions. The most popular intentions aimed at acquiring the information about risk factors, symptoms and signs, and medical care (medicine) of disorders. Very few of the HQs asked for prognosis information about disorders. Many of the HQs (about 10%) asked for general descriptions about a disorder.

4. Location-based and Area-based Feature Weighting

Given an HQ for training (i.e., an HQ with an intention label), both location-based and area-based feature weighting techniques produce feature (word) weights that are used to train the underlying classifier (i.e., SVM). Given an HQ for testing (i.e., an HQ whose intention is to be determined), both weighting techniques are invoked as well to generate feature weights for the classifier. Therefore, the two weighting techniques work as a front-end processor for the underlying classifier in both training and testing.

4.1 Location-based Feature Weighting

As noted in Section 1, a word in an HQ may be more intention-indicative if it appears at the start or the end of the HQ. Therefore, the location-based weighting technique produces two features for each word w in an HQ. The two features are generated based on the distances between w and the start and the end of the HQ, respectively. The larger the distances are, the smaller the feature weights should be.

More specifically, two location weights LW_{front} and LW_{rear} are defined in Equation 1 and Equation 2 respectively. They transform a location p in an HQ q into weights by using the *front* and the *rear* of q as the anchors, respectively.

$$LW_{front}(p,q) = \frac{1}{1 + (\#\, words\ before\ p\ in\ q)} \quad (1)$$

$$LW_{rear}(p,q) = \frac{1}{1 + (\#\, words\ after\ p\ in\ q)} \quad (2)$$

For a location p that is near to the front (rear) of q, its LW_{front} (LW_{rear}) will be large. Both LW_{front} and LW_{rear} of p can be large if q is quite short.

Based on LW_{front} and LW_{rear}, two feature weights $F\ value_{front}$ and $F\ value_{rear}$ for a word w in an HQ q are defined in Equation 3 and Equation 4 respectively.

$$Fvalue_{front}(w,q) = \sum_{p \in \{locations\ of\ w\ in\ q\}} (LW_{front}(p,q)) \quad (3)$$

$$Fvalue_{rear}(w,q) = \sum_{p \in \{locations\ of\ w\ in\ q\}} (LW_{rear}(p,q)) \quad (4)$$

$F\ value_{front}$ ($F\ value_{rear}$) will approach the term frequency (TF) of w only if w appears in those locations that are very near to the front (rear) of q; otherwise $F\ value_{front}$ ($F\ value_{rear}$) of w will be significantly reduced.

It is interesting to note that both $F\ value_{front}$ and $F\ value_{rear}$ are employed as features for the classifier (e.g., SVM) to be more useful in learning a proper model for question classification. As a result, a word w in an HQ has two location-based features (i.e., $F\ value_{front}$ and $F\ value_{rear}$ of w). This is particularly helpful, since category-indicative words for different categories may tend to appear at different locations (start or end) in the HQ. By employing both $F\ value_{front}$ and $F\ value_{rear}$ as the features, the classifier can learn their relative importance using the trainng data. Also note that both $F\ value_{front}$ and $F\ value_{rear}$ of a word w will approach 0 if LW_{front} and LW_{rear} happen to be quite small. In that case, w will have no significant effect. Therefore, the location-based weighting technique expects that only those words that are near to the start or the end of an HQ are category-indicative. It should thus be integrated with the area-based weighting technique, especially when category-indicative words happen to appear far away from the start and the end of an HQ.

4.2 Area-based Feature Weighting

As noted in Section 1, an HQ tends to have an intention c if it has several words that appear in an area of the HQ and are exclusively indicative for c. This area-based weighting technique has two main tasks: (1) identifying the areas (in the HQ) that are related to individual categories, and (2) estimating the feature weight for each word w in the areas by considering how the word is exclusively indicative for specific categories.

To achieve the two tasks, the *correlation* between a word and a category should be defined. We employ the x^2 (chi-square) statistics to estimate the correlation. For a word w and a category c, x^2 (w, c) is estimated by equation 5, where N is the total number of training HQs, A is the

number of training HQs that are in c and contain w, B is the number of training HQs that are not in c but contain w, C is the number of training HQs that are in c but do not contain w, and D is the number of training HQs that are not in c and do not contain w.

$$\chi^2(w,c) = \frac{N \times (A \times D - B \times C)^2}{(A+B) \times (A+C) \times (B+D) \times (C+D)} \tag{5}$$

If $A \times D > B \times C$, w is more likely to appear in c (than in categories other than c) and hence is said to be *positively correlated* to c; otherwise it is *negatively correlated* to c. Equation 6 further defines the *correlation strength* (*CorS*) between w and c.*CorS* (w,c) is in the range of [-1, +1]. It will be positive (negative) if w is positively (negatively) correlated to c.

$$CorS(w,c) = \begin{cases} \dfrac{Log_2(1+\chi^2(w,c))}{Log_2(1+N)}, & \text{if } w \text{ is positively correlated to } c; \\ -1 \times \dfrac{Log_2(1+\chi^2(w,c))}{Log_2(1+N)}, & \text{if } w \text{ is negatively correlated to } c; \\ 0, & \text{otherwise } (w \text{ is not seen in training data}). \end{cases} \tag{6}$$

The algorithm of the area-based feature weighting technique is defined in Figure 3. Given an HQ q, it first identifies the set of areas (in q) that are positively correlated to each individual category c (see Step 1.1 in Figure 3). To identify the positively correlated areas (*PosA*s) for c, the technique sequentially scans the words in q. All words that are negatively correlated to c are skipped, and once a word w that is positively correlated to c is scanned, the starting position of a *PosA* is simply the position of w. The technique then continues to scan subsequent words and maintain the *accumulated correlation strength* (*A-CorS*), which is the sum of the *CorS* values of w and the subsequent words.

Procedure *area-based feature weighting* (q)
Input: An HQ q.
Output: Area-based feature weight of each word in q.
Begin
// **Task I**: For each category c, identify the most positively correlated area ($mPosA_c$) in q
(1) For each category c, do
  (1.1) $POSA_c \leftarrow$ Set of areas (in q) that are positively correlated to c;
  (1.2) $mPosA_c \leftarrow$ The positively correlated area that has the largest *A-CorS* for c;
// **Task II**: Estimate the feature weight of each word in the *mPosA* of each category
(2) For each word w in q, *FeatureWeight*(w) $\leftarrow$ 0;
(3) For p = start position to end position of q, do
  (3.1) $w \leftarrow$ The word at position p in q;
  (3.2) If w is in any $mPosA_c$ (for some category c) and is positively correlated to c, then
    (3.2.1) *MaxAreaStrength*$\leftarrow$ Maximum area strength of w in $POSA_x$, for each category x;
    (3.2.2) *TotalAreaStrength*$\leftarrow$ Total area strength of w in $POSA_x$, for each category x;
    (3.2.3) *FeatureWeight*(w)$\leftarrow$ *FeatureWeight*(w) + (*MaxAreaStrength* / *TotalAreaStrength*);
(4) Return *FeatureWeight*(w) of each word w in q;
End.

Fig. 3. The algorithm for area-based feature weighting, which conducts two tasks: (1) identifying the areas (in the HQ) that are most related to individual categories, and (2) estimating the feature weight for each word w in the areas by considering how the word is exclusively indicative for individual categories.

Figure 4 shows an example to illustrate the idea. The example is for two categories *c1* and *c2* on an HQ with 10 words. Since the first two words are negatively correlated to both categories, they are skipped (and hence the *A-CorS* values for both categories are 0). As the 3rd word is scanned, both categories can have a *PosA* starting at the word, since the word is positively correlated to both categories. The *A-CorS* values for both categories thus start to increase.

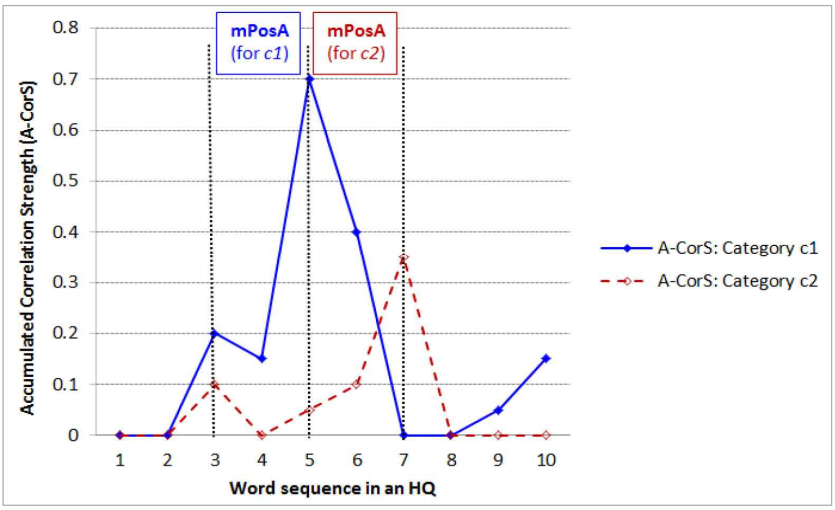

Fig. 4. An example to illustrate the area-based feature weighting technique, which (1) identifies the most-positively correlated area (mPosA) for each category (Category *c1*: the 3rd ~ 5th words; Category *c2*: the 5th ~ 7th words), and then (2) estimates the weight of each word in each mPosA based on the relative strength of the word among all categories.

Therefore a positively (negatively) correlated word will increase (decrease) *A-CorS*. When *A-CorS* becomes less than or equal to 0, there are negatively correlated words whose total *CorS* is compatible with the total *CorS* of positively correlated words. In this case, a*PosA* can be ended at the position at which *A-CorS* is maximized, since such *PosA* is actually dominated by positively correlated words. Following the example illustrated in figure 4, category *c1* has two *PosA*s (the 3rd ~ 5th words; and the 9th ~ 10th words), and category *c1* has two *PosA*s as well (the 3rdword; and the 5th ~ 7th words). Equation 7 is then employed to estimate the *area strength* of each word in a *PosA* of a category *c*.

$$Area\ Strength(w,\ PosA_c) = CorS(w,\ c) \times (\text{maximum } A\text{-}CorS \text{ in } PosA_c) \quad (7)$$

The area strength of a word *w* in a *PosA* for *c* will be amplified if the maximum *A-CorS* in the *PosA* is larger, since in this case *w* has appeared in a nearby area that is more positively correlated to *c*. In the example HQ illustrated in Figure 4, the 5th word *w* is in both $PosA_{c1}$ and $PosA_{c2}$, and hence it has two area strengths $CorS(w,c1)\times0.7$ and $CorS(w,c2)\times0.35$ for *c1* and *c2*

respectively, where 0.7 and 0.35 are the maximum *A-CorS* in $PosA_{c1}$ and $PosA_{c2}$ respectively.

For each category *c*, the technique then selects the "most" positively correlated area (i.e., $mPosA_c$ in Step 1.2 in Figure 3), which is the *PosA* that has the largest *A-CorS* for *c*. In the example shown in Figure 4, $mPosA_{c1}$ spans from the 3rd to the 5th words, while $mPosA_{c2}$ spans from the 5th to the 7th words (see Figure 4).

The technique then generates a feature weight for each word in *mPosA* of each category (see Step 2 and Step 3 in Figure 3). For each word *w* appearing at position *p* in *q*, if *w* is in *mPosA* of a category, the technique estimates its *maximum* area strength and *total* area strength over all categories (see Step 3.2.1 and Step 3.2.2 in Figure 3). In the example shown in Figure 4, the 5th word *w* is in $mPosA_{c1}$ and is positively correlated to *c1*, and hence it will have a feature weight. Its maximum area strength may be $CorS(w,c1)\times0.7$ (for *c1*) and the total area strength may be $CorS(w,c1)\times0.7 + CorS(w,c2)\times0.35$ (for all categories, including *c1* and *c2*). By dividing the maximum area strength by the total area strength, the feature weight for *w* at *p* can be estimated (see Step 3.2.3 in Figure 3). The feature weight actually indicates the *relative* strength of *w* at *p*. It will be larger if *w* at *p* is more *exclusively* and *positively* correlated to certain categories. Each word *w* in an HQ has an area-based feature weight, which is the sum of the feature weights of *w* at different positions in the HQ (see Step 3.2.3 and Step 4 in Figure 3).

5. Experiments

Two experiments (Experiment I and Experiment II) were conducted to evaluate the contributions of the location-based and area-based weighting techniques in (1) classification of HQs and (2) retrieval of relevant HQs, respectively. Table 2 summarizes the main settings for the two experiments.

Table 2. Main settings for the two experiments

	Experiment I	Experiment II
Task	Classification of HQs	Retrieval of relevant HQs
Data source	There were 2,171 Chinese HQs comprehensively collected from 98 sources on the Internet. The intention category of each HQ was determined by a *focused group* approach (see Section 3).	
Data split	5-fold cross validation: The HQs were evenly split into 5 parts, with each part employed as the test data exactly once.	(1) Test data: Randomly selected 50 disorders, and for each disorder selected all HQs about the disorder as the test data (556 test HQs) (2) Training data: All the HQs that were not employed as test data (1,615 training HQs)
Evaluation criteria	(1) Classification F_1 (a) *Micro-averaged* F_1 (b) *Macro-averaged* F_1 (2) *Semantic distance* (SD)	Retrieval F_1 (a) *Micro-averaged* F_1 (b) *Macro-averaged* F_1

5.1 The Data

As noted in Section 3, we have comprehensively collected 2,171 Chinese HQs from 98 sources on the Internet. The HQs were labeled with intention categories. The reliability of the category labeling was ensured by a focused group approach. We thus employed the HQs as the experimental data for both experiments.

In Experiment I, we measured the performance in classifying the HQs. The 2,171 HQs were evenly split into 5 parts so that 5-fold cross validation was conducted - The experiment was conducted 5 times, and in each experiment one part of the data was employed as the test data and the other parts as the training data. Each part was employed as the test data exactly once.

In Experiment II, we measured the performance in retrieving relevant HQs. We randomly selected 50 disorders, and for each of the disorders, extracted all HQs (from the 2,171 HQs) that were about the disorder. For each disorder *d*, we thus had a set (denoted by Q_d) of HQs about *d*. In total we selected 556 HQs, which were employed as the test data, and the remaining 1,615 HQs were employed as the training data. For each disorder *d*, we randomly selected an HQ from Q_d as the *query* HQ, and the remaining HQs as the *candidate* HQs. The query HQ and a candidate HQ were said to be relevant to each other only if they were about the same intention category. The training data was used to train the question classifiers, which then tried to classify both the query HQ and the candidate HQ to determine the relevancy between them. Performance of the classifiers in retrieving relevant HQs was measured. A classifier was poor if it failed to correctly classify the query HQ *or* the candidate HQ.

5.2 Evaluation Criteria

In Experiment I, we employed F_1 and *semantic distance* (SD) as two criteria to evaluate the classification of HQ intentions. F_1 of intention classification was measured by Equation 8, which integrated precision (P_{cat}) and recall (R_{cat}) of HQ classification.

$$\begin{aligned} &\textit{classification } F_1 = \frac{2 \times P_{cat} \times R_{cat}}{P_{cat} + R_{cat}} \\ &P_{cat} = \frac{\textit{Number of correct classifications}}{\textit{Number of classifications made}} \\ &R_{cat} = \frac{\textit{Number of correct classifications}}{\textit{Number of correct classifications that should be made}} \end{aligned} \tag{8}$$

There were two ways to compute average performance in F_1: *micro-averaged F_1* ($MicroF_1$) and *macro-averaged F_1* ($MacroF_1$). In measuring $MicroF_1$, P_{cat} and R_{cat} were computed by viewing all categories as a system, and the resulting P_{cat} and R_{cat} were used to compute $MicroF_1$. $MacroF_1$ is simply measured by averaging F_1 values on *individual* categories.

F_1 evaluated whether the output categories match target categories of the HQs; however, in a hierarchical taxonomy (e.g., the hierarchy of HQ intentions shown in Figure 1), mismatch between

two categories does not necessarily indicate that they are not related (e.g., the "diagnosis" category is related to the "risk factors" category, although they do not match each other). We, therefore, also employed SD (see Equation 9) to measure the "semantic distance" between the two categories.

$$SD(c1,c2)=\begin{cases} Height\ of\ the\ category\ hierarchy, if\ c1\ is\ not\ an\ ancestor\ of\ c2\ and\ vice\ versa \\ Number\ of\ steps\ to\ traverse\ from\ c1\ to\ c2\ in\ the\ hierarchy, otherwise. \end{cases} \quad (9)$$

As an example, in the category hierarchy shown in Figure 1, SD between "diagnosis" and "risk factors" is 1, and SD between "diagnosis" and "homecare" is 3 (i.e., the height of the hierarchy) since they are actually not related to each other. We calculated the total SD on the test HQs. Obviously, a better classifier should achieve a smaller SD.

In Experiment II, we employed F_1 as the criterion as well; however the definitions of precision and recall were modified. F_1 of HQ retrieval was defined in Equation 10, which integrated precision (P_{ret}) and recall (R_{ret}) of HQ retrieval.

$$\begin{aligned} retrieval\ F_1 &= \frac{2\times P_{ret}\times R_{ret}}{P_{ret}+R_{ret}} \\ P_{ret} &= \frac{Number\ of\ relevant\ HQs\ retrieved}{Number\ of\ HQs\ retrieved} \\ R_{ret} &= \frac{Number\ of\ relevant\ HQs\ retrieved}{Number\ of\ relevant\ HQs\ that\ should\ be\ retrieved} \end{aligned} \quad (10)$$

We aimed at measuring how the relevant HQs for each disorder were identified. As in Experiment I, we reported both *micro-averaged F_1* and *macro-averaged F_1*. The former was measured by viewing all the 50 disorders as a system, while the latter was measured by averaging F_1 values on *individual* disorders.

5.3 The Classifiers

We employed SVM as the underlying classification methodology[2], based on two reasons: (1) we focused on investigating how the location-based and area-based features can be used to further improve a learning-based question classifier; and (2) as surveyed in Section 2, SVM is a representative question classifier, because it was routinely employed in question classification and has been shown to be one of the best question classifiers. Therefore, with SVM as the underlying classifier, we can precisely measure the contributions of the location-based and area-based features. Performance of other kinds of features (e.g., syntactic and semantic structures of the HQ) was beyond the scope

2) To implement the SVM classifiers, we employed SVM^{light}(Joachims, 1999), which is publicly available at http://www.cs.cornell.edu/People/tj/svm%5Flight/old/svm_light_v5.00.html. Given an HQ, a distinct word in the HQ corresponded to a feature, with its term frequency (TF) in the HQ as the feature value. Each HQ was thus represented as a feature vector to train and test the SVM classifier.

of our investigation, especially because there is no parser or and analyzer that can derive the syntactic and semantic structure so fall the Chinese HQs (due to the reasons noted in Section 2).

The SVM classifier that employed both the location-based and area-based features was denoted as SVM+Location+Area. Recall that the location-based weighting technique generates two features for each word, while the area-based weighting technique generates one feature for each word. Therefore, in SVM+Location+Area a word corresponded to three features, including two location features and one area feature. To *individually* evaluate the two weighting techniques, we also implemented two classifiers SVM+Location and SVM+Area, which employed the location features and the area features, respectively.

5.4 Experimental Results

Figure 5 shows macro-averaged F_1 of SVM and SVM+Location+Area in intention classification. The result indicated that, by the location features and the area features, the classification was improved in each of the five experimental folds. To verify whether the performance difference was statistically significant, we conducted a significance test, which was a two-sided and paired *t-test*.[3] The result showed that the performance difference was statistically significant (p<0.001), indicating that location-based and area-based feature weightings successfully helped SVM to achieve significantly better performance in intention classification.

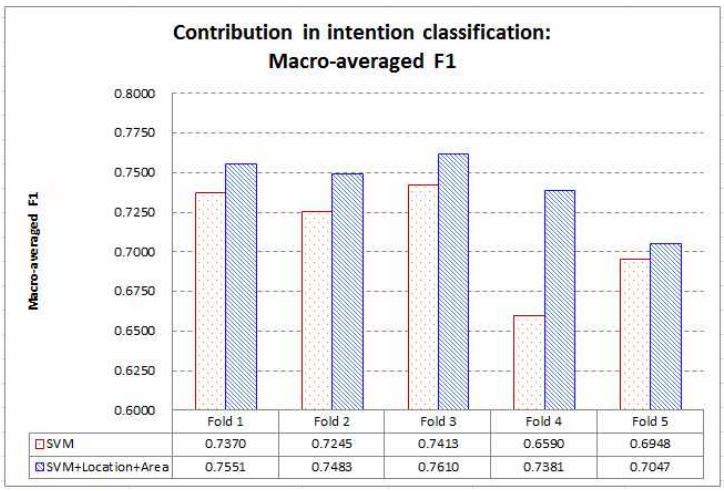

Fig. 5. Contribution in intention classification: With both location-based and area-based features, macro-averaged F_1 of SVM can be significantly improved ($p < 0.001$).

Figure 6 shows micro-averaged F_1 of SVM and SVM+Location+Area in intention classification. The result indicated again that the location features and the area features helped SVM to achieve

3) The t-test was conducted on the F1 values on all the individual categories in the five experimental folds.

better performance in each of the five experimental folds. Since macro-averaged F_1 and micro-averaged F_1 respectively care more about individual categories and "big" categories (those that have many test HQs), the results shown in Figure 5 and Figure 6 together provide strong evidence to justify the contribution of the location-based and area-based feature weighting techniques.

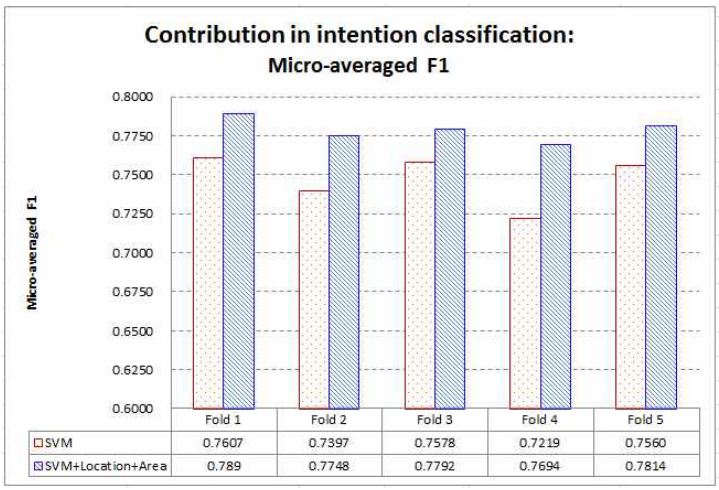

Fig. 6. Contribution in intention classification: With both location-based and area-based features, micro-averaged F_1 of SVM can be improved in each experimental fold.

Figure 7 shows the SD (semantic distance) results. We employed a "most-general" classifier as a baseline, which always assigned the "general description" category (category 1 in Figure 1) to each test HQ. The most-general classifier can represent the most "naive" way of HQ classification. By comparing it with other classifiers,[4] we can measure the contributions of the classifiers in recognizing the intentions of the HQs. We thus calculated the rates of SD reduction (compared with the SD achieved by the most-general classifier) achieved by SVM and SVM+Location+Area (recall that a better classifier should achieve lower SD, ref. Section 5.2, and hence a better classifier should achieve higher SD reduction). The results shown in Figure 7 indicated that both SVM and SVM+Location+Area reduced the SD of the most-general classifier, and hence intention classification with a more intelligent classifier was helpful. It is interesting to note that, SVM+Location+Area achieved a significantly higher SD reduction rate than SVM (p<0.00001). The location features and the area features helped SVM to achieve much lower SD. When comparing SVM and SVM+Location+Area, percentages of the improvement in SD reduction rates ranged from 34.5% (37.84% vs. 50.90% in the 3rd fold) to 84.4% (25.98% vs. 47.90% in the 4th fold). The results reconfirmed the contributions of the location-based and area-based weighting techniques.

4) To make the performance of the classifiers objectively comparable, if a classifier did not classify an HQ into any category, we said that the HQ was classified into the most general category (i.e., the "general description" category) by the classifier. In that case, the classifier actually worked as the "most-general" classifier.

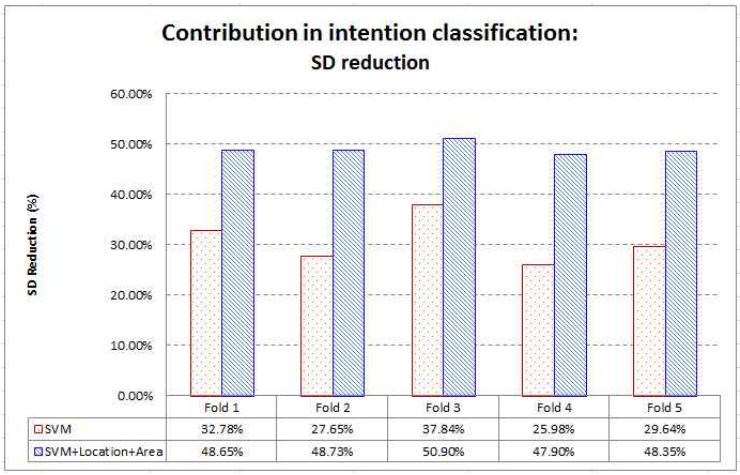

Fig. 7. Contribution in intention classification: With both location-based and area-based features, semantic distance (SD) of SVM can be significantly reduced (p< 0.00001).

We were also concerned with how the location features and the area features *individually* contributed to SVM. Figure 8 shows SD reduction rates of SVM+Location and SVM+Area. We found that the area features alone significantly improved SVM (p<0.001), but the location features did not (p=0.7091). The classifier that employed both types of features (i.e., SVM+Location+area) achieved significantly better performance than the classifiers that employed the location features alone (i.e., SVN+Location, p<0.00001) and the area features alone (i.e., SVM+Area, p<0.001). Both types of features can complement each other in classification of HQ intentions.

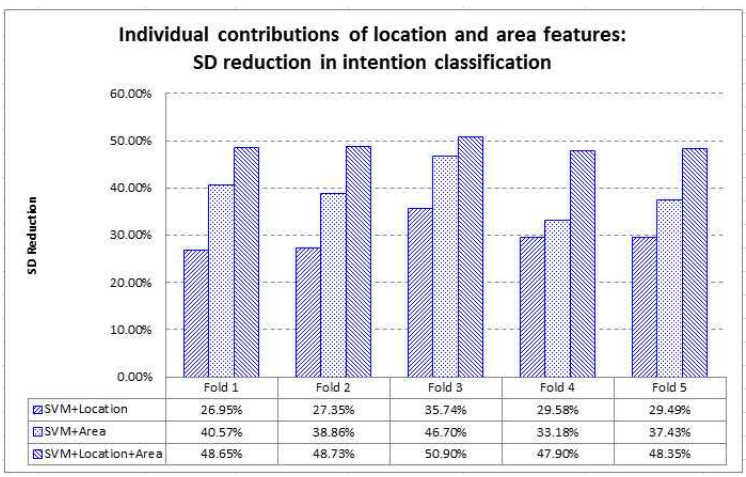

Fig. 8. Contributions of the two types of features (i.e., location-based features and area-based features) in intention classification: The area features can be used to significantly improve SVM (p<0.001), but the location features cannot (p=0.7019). When both types of features are used, performance of the system is significantly better than the two systems that individually employ the two types of features (p<0.0001).

In addition, we were interested in the contributions of the location features and the area features in retrieving relevant HQs. Table 3 shows macro-averaged F_1 and micro-averaged F_1 in retrieving relevant HQs. SVM+Location+Area performed better than SVM in both macro-averaged F_1 and micro-averaged F_1 again. There was 25.73% improvement in macro-averaged F_1 (0.5810 vs. 0.7305) and 15.21% improvement in micro-averaged F_1 (0.7194 vs. 0.8288). Moreover, as we tested 50 query HQs (ref. Section 5.1), we also conducted a two-sided and paired *t-test* on the F_1 values for the 50 queries. The results showed that, the difference between the performance of SVM+Location+Area and SVM was statistically significant (p<0.005). Location-based and area-based feature weightings were thus shown to be helpful in retrieving relevant HQs as well.

Table 3. Performance in retrieving relevant HQs. With the location-based and area-base feature weighting techniques, both macro-averaged F_1 and micro-averaged F_1 in retrieving relevant questions can be significantly improved (p<0.005).

System	Macro-averaged F_1	Micro-averaged F_1
SVM	0.5810	0.7194
SVM+Location+Area	**0.7305**	**0.8288**

6. Conclusion and Future Work

HQs on the Internet are a valuable resource for health promotion and education, since they have been edited (by healthcare professionals) to be both reliable and readable for healthcare consumers. Given a query HQ, retrieval of those HQs that are relevant to the query HQ is thus essential for the utility of the valuable health information. Success of the HQ retrieval heavily depends on the recognition of the intentions of the HQs.

In this paper, we focus on two goals: (1) investigating the intentions of Chinese HQs on the Internet, and (2) developing suitable techniques to improve learning-based intention classifiers. For the former, we analyze over two thousand Chinese HQs from over ninety sources on the Internet, and develop a hierarchy of twelve HQ intentions based on four typical stages of a disorder: *before* the disorder is diagnosed, *when* the disorder is being diagnosed, *after* the disorder is confirmed, and *after* the disorder is treated. We also find that the HQs are not always well-formed and can even be composed of single words, single phrases, or multiple statements. Recognition of the HQ intentions is thus difficult to accomplish by predefining syntactic and semantic rules.

We thus model the intention recognition problem as a text classification problem, and develop two feature weighting techniques to improve a learning-based text classifier, without needing to predefine any rules or patterns. The first feature weighting technique is *location-based*, as it is motivated by the hypothesis that a word in an HQ may be more intention-indicative if it appears at the start or the end of the HQ. The second feature weighting technique is *area-based*, as it is based on the hypothesis that an HQ may have an intention c if it has several words that appear in an area of the HQ and are exclusively indicative for c. Experimental results show that, in both

the classification of HQ intentions and the retrieval of relevant HQs, the two feature weighting techniques can work together to significantly improve SVM, which is one of the best and popular classification methodologies for question classification. The feature weights generated by the location-based weighting and the area-based weighting are thus helpful for the classification of HQs. The contributions are of technical significance to the development of learning-based question classifiers, as well as practical significance to the utility of valuable health information on the Internet. It is interesting to further apply these feature weights to other classification methodologies for various purposes of retrieving health information. It is also interesting to explore the contributions of the location-based and area-based weightings to classifying health questions in different languages.

Acknowledgement

This research was supported by the National Science Council of the Republic of China (Taiwan) under the grant NSC 102-2221-E-320-007.

References

Abbas, J., Schwartz, D. G., & Krause, R. (2010). Emergency medical residents' use of Google® for answering clinical questions in the emergency room. In *Proc. of ASIST 2010*.

Cartright, M.-A., White, R. W., & Horvitz, E. (2011). Intentions and attention in exploratory health search. In P*roceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, 65-74.

Casellas, N., Casanovas, P., Vallbé, J.-J., Poblet, M., Blázquez, M., Contreras, J., López-Cobo, J.-M., & Richard, V. (2007). Semantic enhancement for legal information retrieval: IURISERVICE performance. In *Proceedings of ICAIL*, Palo Alto, CA USA.

Cohen, W. W., & Singer, Y. (1996). Context-sensitive mining methods for text categorization. In *Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval*, 307-315, Zurich, Switzerland.

Eysenbach, G., & Köhler, C. (2002). How do consumers search for and appraise health information on the world wide web? Qualitative study using focus groups, usability tests, and in-depth interviews, *British Medical Journal, 324*, 573-577.

Huang, Z., Thint, M., & Qin, Z. (2008). Question classification using headwords and their hypernyms. Proc. of the 2008 Conference on Empirical Methods in Natural Language Processing, 927-936.

Joachims, T. (1999). Making large-scale SVM learning practical. In *advances in kernel methods - support vector learning*, B. Schölkopf, C. Burges and A. Smola (ed.), MIT-Press.

Krishnan, V., Das, S., & Chakrabarti, S. (2005). Enhanced answer type inference from questions using sequential models. *Proc. of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 315-322.

Lee, C.-W, Day, M.-Y., Sung, C.-L., Lee, Y.-H., Jiang, T.-J., Wu, C.-W., Shih, C.-W., Chen, Y.-R.,

& Hsu, W.-L. (2008). Boosting chinese question answering with two lightweight methods: ABSPs and SCO-QAT. *ACM Trans. Asian Lang. Inform. Process*, Article 12.

Lin, J. & Demner-Fushman, D. (2006). The role of knowledge in conceptual retrieval: A study in the domain of clinical medicine. In *Proceedings of SIGIR' 06*, Seattle, Washington, USA.

Lin, X.-D., Peng, H., & Liu, B. (2006). Support Vector Machines for Text Categorization in Chinese Question Classification. *Proc. of the 2006 IEEE/WIC/ACM International Conferenceon Web Intelligence*.

Liu, R.-L. (2010). Context-based term frequency assessment for text classification. *Journal of the American Society for Information Science and Technology, 61*(2), 300-309.

Liu, R.-L., & Lin, S.-L. (2012). A Conceptual model for retrieval of chinese frequently asked questions in healthcare, *Proc. of the 8th Asia Information Retrieval Societies Conference*, LNCS, Springer-Verlag Berlin Heidelberg, Tianjin, Mainland China.

Liszka, H. A., Steyer, T. E., Hueston, W. J. (2006). Virtual medical care: How are our patients using online health information? *Journal of Community Health, 31*(5), 368-78.

Mishra, M., Mishra, V. K., & Sharma, H. R. (2013). Question classification using Semantic, syntactic and lexical features. *International Journal of Web & Semantic Technology, 4*(3).

Moschitti, A., Chu-Carroll, J., Patwardhan, S., Fan, J., & Riccardi, G. (2011). Using syntactic and semantic structural kernels for classifying definition questions in Jeopardy! *Proc. of the 2011 Conference on Empirical Methods in Natural Language Processing*, 712-724.

Pan, Y., Tang, Y., Lin, L., & Luo, Y. (2008). Question classification with semantic tree kernel. *Proc. of SIGIR'08*, Singapore.

Palotti, J. R. M., Stefanov, V., & Hanbury, A. (2014). User intent behind medical queries: An evaluation of entity mapping approaches with Metamap and Freebase, in *Proceedings of the 5th Information Interaction in Context Symposium*, 283-286.

Peng, F. & Schuurmans, D. (2003). Combining naive bayes and n-Gram language models for text classification. In *Proceedings of 25th European Conference on Information Retrieval Research* (ECIR), 335-350. Pisa, Italy.

Raghavi, K. C., Chinnakotla, M. K., & Shrivastava, M. (2015). "Answer ka type kya he?": learning to classify questions in code-mixed language. in *Proceedings of the 24th International Conference on World Wide Web*, 853-858.

Rekha, V. S., Divya, N., & Bagavathi, P. S. (2014). A hybrid auto-tagging dystem for stackoverflow forum questions. in *Proceedings of the 2014 International Conference on Interdisciplinary Advances in Applied Computing*, Article No.56.

Shuyler, K. S., & Knight, K. M. (2003). What are patients seeking when they turn to the Internet? Qualitative content analysis of questions asked by visitors to an orthopaedics web site, *Journal of Medical Internet Research, 5*(4), 24.

Thomson, M. D., & Hoffman-Goetz, L. (2007). Readability and cultural sensitivity of web-based patient decision aids for cancer screening and treatment: A systematic review, *Medical Informatics and the Internet in Medicine 32*(4), 263-286.

Wang, K., Ming, Z., & Chua, T.-S. (2009). A syntactic tree matching approach to finding similar questions in community-based QA services. In *Proceedings of SIGIR*, Boston, Massachusetts,

USA.

Wu, C.-H., Yeh, J.-F., & Lai, Y.-S. (2006). Semantic segment extraction and matching for Internet FAQ Retrieval. *IEEE Transactions on Knowledge and Data Engineering, 18*(7).

Wu, C.-H., Yeh, J.-F., & Chen, M.-J. (2005). Domain-specific FAQ retrieval using independent aspects. *ACM Transactions on Asian Language Information Processing, 4*(1), 1-17.

Zeng, Q. T., Kogan, S., Plovnick, R. M., Crowell, J., Lacroix, E.-M., & Greenes, R. A. (2004). Positive attitudes and failed queries: an exploration of the conundrums of consumer health information retrieval. *International Journal of Medical Informatics, 73*, 45-55.

Zhang, C., Fan, W., Du, N., & Yu, P. S. (2016). Mining user intentions from medical queries: A neural network based heterogeneous jointly modeling approach, in *Proceedings of the 25th International Conference on World Wide Web* (WWW'16), 1373-1384.

Zhang, D., & Lee, W. S. (2003). Question classification using support vector machines. *Proc. of SIGIR'03*, Toronto, Canada.

[About the authors]

Rey-Long Liu received his Ph.D. from Institute of Computer Science, Tsing Hua University, Taiwan, R.O.C., 1994. He served as a professor and the chairman in Department of Information Management (Chung Hua University) and Department of Medical Informatics (Tzu Chi University). His research lies on intelligent information systems, with special focuses on e-learning and text information systems with applications to biomedicine and business management. He has received many awards in several academic aspects, including teaching, research, and guidance counseling.