Predicting the Uncertainty of Sentiment Adjectives in Indirect Answers

Mitra Mohtarami[1], Hadi Amiri[1,2], Man Lan[3], Chew Lim Tan[1]
[1]SOC, Department of Computer Science, National University of Singapore, Singapore, 117417
[2]NUS Graduate School for Integrative Sciences and Engineering, Singapore, 117456
[3]Institute for Infocomm Research, Singapore, 119613
{mitra, hadi}@comp.nus.edu.sg; lanman@i2r.a-star.edu.sg; tancl@comp.nus.edu.sg

ABSTRACT

Opinion question answering (QA) requires automatic and correct interpretation of an answer relative to its question. However, the ambiguity that often exists in the question-answer pairs causes complexity in interpreting the answers. This paper aims to infer yes/no answers from indirect yes/no question-answer pairs (IQAPs) that are ambiguous due to the presence of *ambiguous sentiment adjectives*. We propose a method to measure the uncertainty of the answer in an IQAP relative to its question. In particular, to infer the *yes* or *no* response from an IQAP, our method employs antonyms, synonyms, word sense disambiguation as well as the semantic association between the sentiment adjectives that appear in the IQAP. Extensive experiments demonstrate the effectiveness of our method over the baseline.

Categories and Subject Descriptors

I.2.7 [**Natural Language Processing**]: Text Analysis

General Terms

Algorithms, Experimentation

Keywords

Sentiment Analysis, Ambiguous Sentiment Adjectives, Indirect yes/no Question-Answer Pairs, Sentiment Ambiguity

1. INTRODUCTION

In indirect yes/no question-answer pairs (IQAPs), the *yes* or *no* words do not explicitly appear in the indirect answers. However, *yes* or *no* responses can be inferred by interpreting the given information in IQAPs. It has been shown that 27% of answers to polar questions do not contain a direct *yes* or *no* word and 44% of them fail to convey a clear *yes* or *no* response [5]. The inherent uncertainty that exists in indirect answers needs to be captured to effectively interpret such answers [8]. It is common that the answerers express their opinion in indirect manner using adjectives with different degree of strength (certainty), e.g. *terrible* has stronger strength than *bad*.

Previous researches showed that adjectives are dominant elements to express opinions [4, 9]. In the review domain, although most of the adjectives have static semantic orientation (SO), positive or negative, the SO of some adjectives vary with context. For example, the adjective *high* has *positive* SO in the phrase *high quality* and *negative* SO in *high cost*. The adjectives with dynamic SO in different contexts are called *Ambiguous Sentiment Adjectives* (ASAs) [10, 1]. Previous works introduced some countable ASAs such as *young*, *many*, *high*, *thick*; they considered other adjectives, like *good* or *terrible*, as unambiguous [10].

In the IQAP domain, we observed that all the ASAs introduced in the review domains can also be ambiguous in this domain. Take the following IQAPs as examples:

E1)	A: Is he ***qualified***?	B: He is ***young***.
E2)	A: Is he ***active***?	B: He is ***young***.

The answers in E1 and E2 contain the ASA *young*. In E1, the answer conveys *no* and in E2 the answer conveys *yes* relative to the adjective used in the questions, i.e. *qualified* in E1 and *active* in E2.

Furthermore, our observation shows that all the adjectives can be potentially ambiguous in the IQAP domain. In the following examples (E3) adapted from [8], the adjective *good* expresses weaker strength than *excellent*, and thus the asker infers that the answerer conveys *yes*:

E3) A: Do you think that's a ***good*** idea, that …?
B: I think it's an ***excellent*** idea.

However, the adjective *good* takes a dynamic certainty with respect to the question and does not convey *yes* all the time, e.g. in E3, if we reverse the adjectives of question and answer then speakers infer that the answer conveys *no* [8]. Thus, the adjective *good* which is always employed to express positive opinions in the review domain, can convey *yes* or *no* in different IQAPs (depending on the adjectives that appears in the question parts). We refer to such adjectives that can be employed in the answers of IQAPs and convey both *yes* and *no* in different answers as ambiguous sentiment adjectives (ASAs) in the IQAP domain.

In this paper, we investigate IQAPs in which polar questions and their corresponding answers contain a sentiment adjective, such as *young*, *good*, *provocative* and *etc*. Therefore, the task is to automatically infer the answer of a given IQAP as *yes* or *no*.

The rest of this paper is organized as follows. Section 2 explains our method for inferring IQAP answers. Section 3 reports the experimental results and evaluation of our method. Section 4 discusses the problems in inferring the answers of IQAPs. The related works are reviewed in Section 5. Finally, Section 6 concludes the paper.

2. METHOD

Our method has four main stages to infer the *yes* or *no* answers to IQAPs. First, we measure the certainty of the answer relative to its question for all IQAPs. Second, for each IQAP, we compute a threshold to evaluate the certainty of answer toward *yes* or *no*

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.
Copyright 2011 ACM 978-1-4503-0717-8/11/10...$10.00.

responses. Third, we infer the answers in each IQAP using the certainty of its answer and its obtained threshold. Finally, we present a refinement on the method by using synonyms. We explain these stages in the subsequent sections respectively.

2.1 Assigning Degree of Certainty to Answers

In this section, we aim to compute the certainty of an answer relative to its question in a given IQAP. Such certainty can be computed based on the association between the adjective of the question (SAQ) and the adjective of the answer (SAA). If the association between the SAQ and SAA is high, then the certainty of the answer relative to its question will be high.

Any similarity measure can be employed to estimate the association between SAQ and SAA. We here use two popular measures, Pointwise Mutual Information (PMI) and Latent Semantic Analysis (LSA) for this purpose. PMI between two words measures the mutual dependence of them and is defined as follows [9]:

$$PMI(w_1, w_2) = \log_2 \left(\frac{P(w_1, w_2)}{P(w_1)\, P(w_2)}\right) \qquad (1)$$

where $P(w_1, w_2)$ is the probability that w_1 and w_2 co-occur in the same context (e.g., a fixed window or sentence), and $P(w_1)$ and $P(w_2)$ are the probability of w_1 and w_2 in the entire corpus. Since PMI requires a large corpus to be effective, in our experiments, we employ a large corpus of reviews explained in Section 3 to calculate PMI between the words, and consider co-occurrence of two words in less than five words distance between them.

LSA is another learning method for computing similarity between words [6]. It works based on analyzing relationships between a document and the words that it contains. LSA performs several steps to compute the association between two words. First, it forms a matrix with documents as rows and words as columns. Cells contain the number of times that a given word is used in a given document. Second, it employs Singular Value Decomposition (SVD) to represent the words and documents as vectors in a high dimensional semantic space. In a new matrix, words are represented as vectors. Finally, similarity of two words is computed as the cosine between their corresponding vectors in the semantic space. The value of cosine will be +1 for identical meanings, zero for unrelated meanings and -1 for opposite meanings.

We obtained LSA values using the TASA corpus[1] instead of the corpus that we employed for PMI. This is because LSA is computationally expensive with large corpora. In fact, this is because LSA uses a word-by-document matrix and the cost of computation increases substantially with very large corpus.

2.2 Defining a Threshold

As we discussed above, we need to know whether the answer in an IQAP has enough certainty to convey a *yes* or *no* response. We compute a threshold for this purpose which can vary in different IQAPs. Since the antonym of a word belongs to the same scalar group (e.g., *hot* and *cold*) and has different semantic orientation with the word [4], we can utilize the antonym of the SAQ to compute a threshold for each IQAP. Our intuition is that if the association strength between an SAA and its corresponding SAQ is greater than the association strength between the SAA and the antonym of its SAQ, the answer has enough degree of certainty to convey *yes*, and otherwise the answer is more likely to be uncertain relative to the question and conveys a *no* response. For example, in E1, the association strength between *young* and *qualified* is smaller than the association between *young* and *unqualified*; therefore, the answer conveys *no*.

We find the antonym of an SAQ in two steps. First, we employ the IMS word sense disambiguation system [11] to detect the sense of the SAQ. Then, we use WordNet to get the antonym of the SAQ based on its predicted sense. In WordNet, different senses of a word can have different antonyms.

2.3 Inferring Yes or No Answers

The following decision procedure employs two previous steps to decide what a given answer conveys:

$$answer = \begin{cases} yes, & assoc(SAQ, SAA) > assoc(\sim SAQ, SAA) \\ no, & assoc(SAQ, SAA) < assoc(\sim SAQ, SAA) \\ uncertain, & otherwise \end{cases} \qquad (2)$$

where $assoc(.,.)$ indicates our similarity measure (either *PMI* or *LSA*), and $\sim SAQ$ is antonym of the SAQ. Note that the appearance of a negation word in the answer to a question reverses the inferred answer, thus here it flips *yes* and *no* responses, but *uncertain* remains unchanged.

2.4 Refining Using Synset

In this section we propose to use the synonyms of the SAAs to supplement our method with more information about the SAAs. Since the synonym of a word is a word that has the same or nearly the same meaning as the original word, the SAA can be replaced by any of its synonyms with no major changes in its inferred original answer. In addition, different senses of an SAA may have different sets of synonyms (synsets) in WordNet. We can obtain the synset of SAAs using a word sense disambiguation system and WordNet in a similar way that we did for antonyms in Section 2.2. Having the synset of an SAA, we compute the association between SAQ and the synset of the SAA as follows:

$$assoc(SAQ, syn(SAA)\,) = \frac{1}{|synset|} \sum_{i=1}^{|synset|} assoc(SAQ, syn_i(SAA)) \qquad (3)$$

where $syn(SAA)$ is the synset of the SAA, and $\mathrm{syn_i(SAA)}$ is the i^{th} word in $syn(SAA)$. Equation 3 computes the association between SAQ and the synset of SAA by averaging the sum of the association between SAQ and each of the synonyms for SAA.

As we discussed before, the antonym of an SAQ can be used to decide about the certainty of the answer for an IQAP as *yes* or *no*. In Section 2.3 the antonym has been used with the SAA itself, i.e. $assoc(\sim SAQ, SAA)$. Here we use the synset of the SAA to predict the association between $\sim SAQ$ and the SAA more precisely. The following Equation can be used for this purpose:

$$assoc(\sim SAQ, syn(SAA)\,) = \frac{1}{|synset|} \sum_{i=1}^{|synset|} assoc(\sim SAQ, syn_i(SAA)) \qquad (4)$$

We can use the above two equations to infer the yes or no response as follows:

$$answer = \begin{cases} yes, & assoc\big(SAQ, syn(SAA)\big) > assoc(\sim SAQ, syn(SAA)) \\ no, & assoc\big(SAQ, syn(SAA)\big) < assoc(\sim SAQ, syn(SAA)) \\ uncertain, & otherwise \end{cases} \qquad (5)$$

[1] LSA is obtained from: http://cwl-projects.cogsci.rpi.edu/msr

3. EVALUATION AND RESULTS

In this section we first explain the datasets that we used in this research, and then report the experiments conducted to evaluate our approach.

We used the dataset developed in [8] to evaluate our method. This dataset contains a set of IQAPs and their corresponding *yes* or *no* labels as its ground truth. It includes 125 IQAPs with two different sentiment adjectives in any question-answer pair as described in [8]. They used two sources to gather the IQAPs: five different shows from online CNN interview transcripts, and the Switchboard Dialog Act corpus. They manually annotated the IQAP dataset for *yes* or *no* responses and identified the adjectives of the questions and answers. In all instances of IQAPs, the SAA is different from the SAQ.

We also used two datasets as development datasets to compute the association strength of word pairs based on PMI and LSA measures. To compute the co-occurrence information for PMI, we collected a large corpus of 1.5M reviews from Amazon product reviews for 25 different product types, such as *book*, *video*, and *music*. However, as we discussed in Section 2.1, LSA in contrast to PMI cannot handle large corpora [7]. Therefore we employed the standard Touchstone Applied Science Associates (TASA) corpus to compute the association strength of word pairs using LSA. The TASA corpus is a collection of texts from textbooks, literature, works of fiction and nonfiction used in schools and the reading materials that a person is supposed to have been exposed to by his first year in college. This corpus contains more than 17M tokens corresponding to around 155K different types.

3.1 Experimental Results

In this section we report detail results of our approach with different configurations. We compare the approach proposed in [8] as a baseline.

Given an IQAP, [8] assigned a semantic orientation (SO, referred as expected rating value by authors) to both SAA and SAQ of the given IQAP, and then interpreted the answer based on the SOs. For instance, if the SOs of the SAA and SAQ have different signs, then the answer conveys *no*. In case of the same sign, if the SO of the SAA is greater than or equals to the SO of the SAQ, then the answer conveys *yes*, and otherwise *no*. They obtained an accuracy of 60% on the same IQAP dataset which is used in our experiments. They used an external source (a large corpus of reviews with rates) to compute the SO of adjectives. Given an adjective, they computed the SO of the adjective as a function of the probability of rate given the adjective.

Their approach assigns a globally fixed SO score to each adjective. For example, the adjectives *best* and *great* are assigned the fixed SO scores of 1.08 and 1.1 respectively. This approach ignores the context in which the adjectives appear (i.e. the IQAP). However, in our approach the degree of certainty for the same answer may change in different IQAPs. This dynamic degree of certainty not only depends on the SAA itself but also on the SAQ that appears in the IQAP. So, our method utilizes the context information better than the method proposed in [8].

Table 1 shows the results of different approaches in terms of precision, recall and f-measure. The results in Table 1 are based on Equation 5 where we use antonyms (of SAQs) and synsets (of SAAs) to infer the yes or no answers. The first and second rows of Table 1 shows the result of our method when it uses PMI and LSA respectively. The last row shows the baseline results.

Table 1. Performance of different approaches

Similarity Measures	Precision	Recall	F-Measure
PMI-synset-antonym	67.14	65.45	66.28
LSA-synset-antonym	**73.97**	**75.98**	**74.96**
Baseline [8]	60.00	60.00	60.00

As it is clear from Table 1, when we use LSA our method achieves better performance than PMI. It was expected since PMI is known as a contextual similarity measure while LSA is known as a semantic similarity measure. So, LSA can better measure the semantic association between the adjectives which can definitely help the inference process. Our method using both PMI and LSA significantly outperforms the baseline method.

4. ANALYSIS AND DISCUSSION

In this section, we discuss the effectiveness of our method from different perspectives. As we mentioned before, our method utilizes synset, antonym, and word sense disambiguation techniques. In this section, we dig into the IQAP problem and investigate the effectiveness of these techniques to tackle the IQAP problem. In Section 4.1, we analyze the role of synsets and antonyms, while in Section 4.2 we discuss the role of WSD.

4.1 Role of Synsets and Antonyms

In this section, we evaluate the effectiveness of synsets and antonyms for inferring yes or no answers. For this purpose, we repeat the experiments by ignoring synsets or antonyms respectively. Table 2 shows the results.

In Table 2, PMI-Antonym (LSA-Antonym) shows the results when we use Equation 2 to infer the answer of an IQAP based on the PMI (LSA) measure. In Equation 2, we only use SAA, SAQ, and ~SAQ and do not utilize the synsets (of SAAs). Table 2 shows that ignoring the synsets results in significant reduction in the final performance, i.e. from 66.28% (see Table 1) to 57.53% for PMI and 74.96% to 61.55% for LSA.

This result shows that synsets are highly effective for IQAP problem. We believe one of the reasons is about the fact that some SAAs and SAQs never (or rarely) co-occurred in our large corpus. This results in a very low association between them. However the synonyms of the SAAs may frequently occur with the SAQs. Therefore, the synonyms help us to more reliably predict the association between the SAQs and SAAs and consequently better infer the yes or no responses. Similar to PMI, LSA can benefit from synsets. In fact, as the result shows, LSA benefits more from the synsets than PMI. The reason is that our LSA measure uses a smaller corpus (TASA) than PMI. Therefore, it is more likely that an SAA do not appear in the LSA corpus than PMI corpus. In that sense, LSA should benefit more from the synsets than PMI.

Table 2. Performance without using synset

Similarity Measures	Precision	Recall	F-Measure
PMI-antonym	58.94	56.18	57.53
LSA-antonym	62.23	60.88	61.55
PMI-synset	34.86	41.69	37.97
LSA-synset	67.35	56.86	61.66
PMI	32.20	34.38	33.25
LSA	66.70	54.95	60.26

To investigate the role of antonyms, we repeat the experiments without using them. In other words, for each IQAP, the answer is interpreted only based on the association between the SAQ and the synset of the SAA. Given an IQAP, if the similarity association between SAQ and the synset of the SAA is positive, then the inferred answer will be *yes*; if it is negative, the inferred answer will be *no*, and otherwise, it will be *uncertain*. The results of these experiments are shown as PMI-synset and LSA-synset in Table 2 for PMI and LSA respectively.

As it is clear from Table 2, the antonyms can also help to infer the correct answer comparing to PMI-synset-antonym or LSA-synset-antonym (See Table 1). The results of both PMI and LSA have significantly decreased when we do not use antonyms, from 66.28% to 37.97% for PMI and 74.96% to 61.66% for LSA. In addition, it is notable that the performance of PMI decreased more than LSA (from 66.28% to 37.97%).

Finally, we apply the proposed method with no using synsets and antonyms. Here, for each IQAP, the answer is interpreted only based on the association between the SAQ and the SAA. Given an IQAP, if the similarity association between SAQ and the SAA is positive, then the inferred answer will be *yes*; if it is negative, the inferred answer will be *no*, and otherwise, it will be *uncertain*. The results of these experiments are shown in the last two rows of the Table 2. As expected, we see the lowest performance when we do not utilize both synonyms and antonyms.

4.2 Role of Word Sense Disambiguation

In our method, we employed an automatic WSD system and obtained 66.28% and 74.96% performance using PMI and LSA respectively (see Table 1). Here, we study the impact of the WSD system on these results.

For this purpose, instead of the WSD system we only used the most common sense of the adjectives (the first sense in WordNet) and repeat the experiments. We took the most common sense as a replacement for the WSD system because it has been shown as a strong baseline in the WSD area. The results are shown in Table 3. As it is clear, using the most common sense of the adjectives slightly reduces the performance, from 66.28% to 65.97% for PMI and 74.96% to 73.83% for LSA.

Table 3. Performance without using IMS

Method	Precision	Recall	F-Measure
PMI	66.35	65.59	65.97
LSA	73.02	74.66	73.83

It is notable that, in this experiment, the WSD system has assigned the most common sense to around 80% of the adjectives. In other word, only 20% of the adjectives assigned senses different than their most common senses. The efficiency of the WSD could have been more highlighted, if more IQAPs contain adjectives with senses different from their most common senses.

5. RELATED WORK

In [3] the authors presented a computational model for interpreting and generating indirect answers to polar (Yes/No) questions using a discourse-plan-based approach and a hybrid reasoning model. [8] worked on indirect yes/no question-answer pairs involving an adjective in question and an adjective in the answer. As we explained in Section 3.1, they used an external source to assign a SO (or ER in [8]) to each adjective (SAA and SAQ) to do this task. For each IQAP, if the SOs of the SAA and SAQ have different signs, then the answer conveys *no*. In case of the same sign, if the SO score of the SAA is greater than or equals to the SO of the SAQ, then they inferred a *yes* response; if the SO score of the SAA is smaller than the SO of its SAQ, then they inferred a *no* response. They also used the method proposed in [2] to compute the SO scores using WordNet instead of the external source. They showed that using WordNet produces 56% performance for inferring *yes* or *no* answers to IQAPs. Although they assigned a static SO to each adjective, ASAs have dynamic SO. Previous works on ASAs considered the adjectives in review domain [10, 1]. However, in this paper we examined the behavior of adjectives in the IQAP domain. We showed that the adjectives can show different degrees of certainty in different question-answer pairs which can lead to different responses.

6. CONCLUSION

In this paper we examine the behavior of adjectives in Indirect yes/no Question-Answer Pairs (IQAPs) domain. In particular, our task is to automatically detect whether the answer of a given IQAP conveys *yes* or *no*. We show that measuring the association between the adjectives in question and answer can be a main factor to infer a clear response from an IQAP. We utilize antonyms, synonyms and word sense disambiguation to tackle the IQAP problem and investigate the effectiveness of each of these techniques for this task.

7. REFERENCES

[1] Balahur, A. and Montoyo, A. 2010. OpAL: Applying Opinion Mining Techniques for the Disambiguation of Sentiment Ambiguous Adjectives. *ACL*, pp. 444-447.

[2] Blair-Goldensohn, S., Hannan, K., McDonald, R., Neylon, T., Reis, G., and Reynar, J. 2008. Building a sentiment summarizer for local service reviews. *NLPIX.*

[3] Green, N. and Carberry, S. 1999. Interpreting and generating indirect answers. *Computational Linguistics*, 25(3), 389-435.

[4] Hatzivassiloglou, V. and McKeown, K. R. 1997. Predicting the semantic orientation of adjectives. *ACL*, pp. 174-181.

[5] Hockey, B., Knill, D., Spejewski, B., Stone, M., and Isard, S. 1997. Can you predict answers to Y/N questions? Yes, No and Stuff. In *Proceedings of Eurospeech*, pp. 2267-2270.

[6] Landauer, T., Foltz, P., and Laham, D. 1998. An introduction to Latent Semantic Analysis. *Discourse Processes*, 259-284.

[7] Lindsey, R., Veksler, V., Grintsvayg, A., and Gray, W. 2007. Be Wary of What Your Computer Reads: The Effects of Corpus Selection on Measuring Semantic Relatedness. *ICCM*, pp. 279-284.

[8] Marneffe, M. C., Manning, C. D., and Potts, C. 2010. "Was it good? It was provocative." Learning the meaning of scalar adjectives. *ACL*, pp. 167–176.

[9] Turney, P. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *ACL*, pp. 417-424.

[10] Wu, Y. and Jin, P. 2010. Disambiguating Sentiment Ambiguous Adjectives. *SemEval-ACL*, pp. 81-85.

[11] Zhong, Z. and Ng, H. T. 2010. It Makes Sense: A Wide-Coverage Word Sense Disambiguation System for Free Text. *ACL*, pp: 78–83.