Arabic Phrase-Level Contextual Polarity Recognition to Enhance Sentiment Arabic Lexical Semantic Database Generation

Samir E. Abdelrahman, Hanaa Mobarz, Ibrahim Farag
Computer Science Department
Faculty of Computers and Information
Cairo, Egypt

Mohsen Rashwan
Electronics and Communications Department
Faculty of Engineering
Cairo-Egypt

Abstract—Most of opinion mining works need lexical resources for opinion which recognize the polarity of words (positive/negative) regardless their contexts which called prior polarity. The word prior polarity may be changed when it is considered in its contexts, for example, positive words may be used in phrases expressing negative sentiments, or vice versa. In this paper, we aim at generating sentiment Arabic lexical semantic database having the word prior coupled with its contextual polarities and the related phrases. To do that, we study first the prior polarity effects of each word using our Sentiment Arabic Lexical Semantic Database on the sentence-level subjectivity and Support Vector Machine classifier. We then use the seminal English two-step contextual polarity phrase-level recognition approach to enhance word polarities within its contexts. Our results achieve significant improvement over baselines.

Keywords—Sentiment Arabic Lexical Semantic Database; Support Vector Machine; Contextual Polarity

I. INTRODUCTION

Opinion mining is the task to distinguish between subjective and objective sentiments in the text. Most work of opinion mining has been extensively explored at document-level while there has been few researches investigating feature design at the sentence-level. Any sentence may have positive, negative and neutral opinions, for example, "I have been working hard and over the past few months but the results were bad" and it is difficult to accurately mark subjective phrase boundaries such that the polarity classification may differ substantially from the sentence-level and the document-level in that resulting bag-of-words feature vectors tend to be very sparse resulting in lower classification accuracy [1].

General approach of opinion mining is to start with database having positive and negative word with their prior polarities, i.e. the initial word polarities regardless their contexts. For example, ["good", "happiness", "wonderful"] have positive prior polarities and ["bad", "hateful", "sadness"] have negative prior polarities. However, contextual polarity of the phrase in which a word appears may be different from the word's prior polarity. As in the following example:-

The above words have positive prior polarities, but they are not all being used to express positive sentiments. For example, ["work"] is preceded by a negative tool ["لم"] ["not"] so it has a negative contextual polarity. Also, the words ["organization", "security", "cooperation", "organization"] have neutral contextual polarities because they are organization names. But the words ["terrorism", "war", "objection", "comply"] have similar prior and contextual polarities. Also, these words ["terrorism", "war", "objection", "comply"] have negative prior polarities but they are not all being used to express negative sentiments. For example, the expression ["war against terrorism"] gives positive sentiment and the rest of words have similar prior and contextual polarities.

There are many things should be taken into consideration in the phrase-level contextual recognition. Negation may reverse the prior polarity of the term. It may precede the term directly ["I am not good"] or it may involve long distance dependency such as ["Western civilization can't configure a global system happier"]. Intensifiers influence the force of the term ["very", "deeply", "too", ...]. Shifter words precede or follow the polar term and influence its polarity, for example, ["prevent punishment"]. Connectors

www.ijacsa.thesai.org
also may influence the contextual polarity; there are some connectors give similar polarities for all connected words ("and", "but") and some connectors extract different polarities ("but", "or") and some connectors extract different polarities ["otherwise", "besides", "but"] ["On the contrary", "contrast", "but", "in spite of" ...].

We used SentiRDI [2] which is a large set of subjective clues coupled with their prior polarities; subjective clues are words with polar (positive/negative) prior polarities. We considered each phrase having one of these clues to classify its contextual polarity. To classify the contextual polarities, we used the seminal English work approach [3] that first determines if the phrases are polar or neutral and then it takes the polar phrases for additional classification to determine the polarity for each polar phrase. In our research, all annotations and classification results were manually revised and assessed. For the classification assessment, we used F-measure (F), Precision (P), and Recall (R).

This paper is organized as follow: Section II describes in brief some main contextual polarity related works. Section III gives the overview of prior polarity subjectivity Arabic database (SentiRDI). Section IV describes the corpus that is used in sentence subjectivity classifier and contextual polarity. Section V describes the sentence subjectivity classification using Support Vector Machine (SVM). Section VI explains the contextual polarity influencers and proposed features that are used in the two-step phrase-level classification approach [3]. Section VII shows the experimental results of contextual polarity. Section VIII shows the analysis of the experimental results. Finally, Section IX draws our conclusions and future work.

II. CONTEXTUAL POLARITY RELATED WORK

Nowadays, many researches have been contributing to the contextual polarity recognition task at various textual levels such as [1, 3, 4, 5]. They mainly classified expressions related to some subjective clues. Also, they often used manual developed lexicons to help in classifying polarities. Per to our knowledge, there is no robust and tested phrase-level contextual polarity study in Arabic.

III. PRIOR POLARITY SUBJECTIVITY DATABASE

Our approach uses an Arabic lexical Resource for opinion mining (SentiRDI) [2] which has the subjectivity and the orientation of more than 18,400 semantic fields covering over 150,000 words in Arabic. Subjective semantic fields in the database are the subjective clues [1, 3] which are words used to express private states [6] mainly an opinion, emotion, evaluation, stance, speculation etc.

IV. RESEARCH CORPUS

We translated MPQA opinion corpus in Arabic which consists of 535 English-language news articles from a variety of sources, manually annotated [7] for subjectivity analysis. The corpus consists of 9700 sentences, 55% of them are labeled as subjective, while the rest are objective. We consider only 3578 sentences with 18,678 subjective phrases. Subjective phrase is the expression which contains subjective clue (term that has subjective prior polarity). The translated annotations were manually revised and corrected by all authors.

V. SUBJECTIVITY CLASSIFICATION

Simple text preprocessing was executed in order to remove special characters and non-Arabic characters in corpus. More advanced text preprocessing was executed in order to prepare it for SVM algorithm input such as extracting named entities using [8], assigning Part Of Speech tags (POS) using the Research and Development International (RDI) and assigning the prior polarity of each word by using SentiRDI. The features that were extracted from the sentence are:

The word Part of Speech (POS): RDI-ArabMorphoPOS tagger was used [9].

We used our prior polarity semantic database (SentiRDI) to determine the polarity of each word to acquire the following four features: Number of positive noun; Number of positive verb; Number of negative noun; Number of negative verb.

Average Polarity of sentence = \( \frac{1}{n} \sum_{i=1}^{n} P_{wi} \) (1)

Where \( n \) is number of words in sentence, \( P_{wi} \) polarity of word \( i \) in sentence that is specified before from prior polarity database (SentiRDI) such that

\[ P_{wi} = \begin{cases} -1 & \text{negative} \\ 0 & \text{objective} \\ 1 & \text{positive} \end{cases} \] (2)

Average Term Frequency: Inverse Sentence Frequency (TF-ISF) for sentence (Si) can be computed by the following equation:

\[ \text{Avg TF-ISF}_{Si} = \frac{1}{|Si|} \sum_{t=1}^{|Si|} (TF_{t,si} \times ISF_t) \] (3)

Where TF presents the number of occurrences of each term within the sentence and can be normalized by dividing it by size of sentence.

\[ TF_{t,si} = \frac{N_{t,si}}{|Si|} \] (4)

Where \( N_{t,si} \) is the number of occurrences of term \( t \) in sentence \( S \). \( |Si| \) is the number of words in sentence \( S \). ISF is used for terms that appear in the small number of sentences. This factor is useful because numbers of subjective terms are small compared with neutral (objective) ones.

\[ ISF = \log \frac{S}{S_i} \] (5)

Where \( S \) is the number of all sentences in the corpus and \( S_i \) is the number of sentences containing term \( i \).

The results of SVM are 77.7%, 75.01%, and 80.6% for F-measure, Precision, and Recall respectively.

VI. CONTEXTUAL POLARITY

A. Contextual polarity influencers

There are a lot of factors [3] that influence the prior polarity of term:–

1 http://mpqa.cs.pitt.edu/

2 http://www.rdieg.com/
Negation: it is considered one of the most factors that influence the contextual polarities of subjective clues. Negation reverses the prior polarities of the subjective clues which may be local. For example, one of the Arabic negation tool may precede the subjective clue directly ["لا، فرنسا لن تتفاهم على هذه الصيغة"] ["France will not agree to this formula"] or it may have long distance dependency of the clue as ["لا يوجد لجنة مساعدة لولاية " which translates as "the committee on the CFP of the country."

Inflection: (Neutral contextual polarities) clues are words that may be local or non-local. It may be local if the subjective clue is one of objective shifters or not. The features set used in this classifier are:  

1. **Prior polarity of subjective clue:** It is the prior polarity of the subjective clue, it is a binary value (0 if it is positive or 1 if it is negative).
2. **Semantic ID of SC:** It is the feature presents the RDIArabSemanticDB word semantic field identification. This feature is designed to help in recognizing the meaning of SC decreasing the ambiguity of the word sense.
3. **POS of SC:** It is the POS that presents POS tag of the SC previous word.
4. **POS of next word:** It is the POS that presents POS tag of the SC next word.
5. **Prior polarity of SC:** It is the prior polarity of the subjective clue from SentiRDI. This feature has a binary value of (0) if it is positive or (1) if it is negative.
6. **NER_SC:** It is the binary feature to present if the subjective clue is a named entity.
7. **SC_before:** It is the binary feature to present if the subjective clue is preceded by another one.
8. **SC_after:** It is the binary feature to present if the subjective clue is followed by another one.
9. **Self_intensifier:** It is the binary feature to present if subjective clue is one of intensifiers or not.
10. **Intensifier_before_after:** It is the binary feature to present if there is intensifier before or after the subjective clue.
11. **Connector:** It is the binary feature to present if there is a connector ["و", "و"] ["and", "or"] between two subjective clues (in this case they have the same polarity).
12. **Shift_conn:** It is the binary feature to present if there is a connector ["لكن", "بالمثل من"] ["but", "in spite of"] between two subjective clues (in this case they have opposite polarity).
13. **Obj_shifter:** It is the binary feature to present if there is one of objective shifters before a subjective clue.
14. **Self_obj_shifter:** It is the binary feature to present if the subjective clue is one of objective shifters or not.

---

### Table I. Baseline Classifier Results

<table>
<thead>
<tr>
<th></th>
<th>Positive Expression</th>
<th>Negative Expression</th>
<th>all</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>67.6</td>
<td>42.6</td>
<td>52.4</td>
</tr>
<tr>
<td>P</td>
<td>76.6</td>
<td>35</td>
<td>48.45</td>
</tr>
<tr>
<td>R</td>
<td>60.1</td>
<td>54.4</td>
<td>57.1</td>
</tr>
</tbody>
</table>
D. Features of Polarity classification

This is the second-step classifier that takes all polar expressions are produced from the first-step neutral-polar classifier to determine whether the contextual polarity is positive or negative. The features set used in this classifier are:

- **Word:** it is the word which has non-neutral prior polarity subjective clue (SC).

- **Semantic ID of SC:** it presents the RDIArabSemanticDB semantic field identification which helps in recognizing the meaning of the subjective clue decreasing the ambiguity of word sense.

- **Prior polarity of SC:** it is the prior polarity of subjective clue extracted from SentiRDI. This feature has a binary value that takes value (0) if it is positive or (1) if it is negative.

- **Prior polarity of next word:** it presents the prior polarity of the SC next word.

- **Prior polarity of previous word:** it presents the prior polarity of the SC previous word.

- **Self_intensifier:** it is the binary feature to present if the SC is one of intensifiers or not.

- **Intensifier-before-after:** it is the binary feature to present if there is an intensifier before or after the subjective clue.

- **Connector:** it is the binary feature to present if there is a connector ["و", "أ" ] ["and", “or"] between two subjective clues (in this case they have similar polarities).

- **Shift_conn:** it is the binary feature to present if there is a connector ["لكن", "بالنِّتِمَة" ] ["but", "in spite of"] between two subjective clues (in this case they have opposite polarities).

- **Negation:** it is the binary feature to present if the subjective clue is preceded by one of the shifter; these shifter alter the polarity to its opposite.

- **Negative polarity shifter:** it is the binary feature to present if the subjective clue is preceded by one of the shifter; these shifter alter the polarity to its negation.

- **Positive polarity shifter:** it is the binary feature to present if the subjective clue is preceded by one of the shifter; these shifter change polarity to its affirmative.

VII. EXPERIMENTAL RESULTS OF CONTEXTUAL POLARITY

The objective of the experiments is to classify the contextual polarities of the expressions that contain instances of the subjectivity clues from SentiRDI. Support vector machine (SVM) is used for the classification task. In order to classify the contextual polarities of subjective expressions, first we determine whether clue instances are neutral or polar in context (the results of this classifier shown in Table II). Second, all the polar clues that result from the first-step are considered for more classification to determine whether the polar clue instance is positive or negative polar polarity (the results of this classifier shown in Table III).

**TABLE II. STEP 1 SVM CLASSIFIER RESULTS**

<table>
<thead>
<tr>
<th></th>
<th>WT</th>
<th>WT + PP</th>
<th>All</th>
<th>WT</th>
<th>WT + PP</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Polar</td>
<td>F</td>
<td>78.4</td>
<td>87.6</td>
<td>91.5</td>
<td>72</td>
<td>81.8</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>77</td>
<td>80</td>
<td>86.7</td>
<td>60.5</td>
<td>69.4</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>80</td>
<td>96.8</td>
<td>96.8</td>
<td>89</td>
<td>99.8</td>
</tr>
</tbody>
</table>

Table II presents the results of neutral-polar classifier for the 15-feature classifier and two baseline classifiers. Table III presents the results of polarity classifier for the 13-feature classifier and two baseline classifiers. The two baseline classifiers are the word token (WT) classifier and the word token with prior polarity (WT + PP) classifier.

**TABLE III. STEP 2 SVM CLASSIFIER RESULTS**

<table>
<thead>
<tr>
<th></th>
<th>WT</th>
<th>WT + PP</th>
<th>All</th>
<th>WT</th>
<th>WT + PP</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>F</td>
<td>79.2</td>
<td>81.2</td>
<td>81.3</td>
<td>76.2</td>
<td>81</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>75.7</td>
<td>80.7</td>
<td>80.8</td>
<td>80.2</td>
<td>80.7</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>83</td>
<td>81.8</td>
<td>81.9</td>
<td>72.6</td>
<td>81.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>WT</th>
<th>WT + PP</th>
<th>All</th>
<th>WT</th>
<th>WT + PP</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Negative</td>
<td>F</td>
<td>83.1</td>
<td>81.4</td>
<td>81.6</td>
<td>83.1</td>
<td>83.1</td>
</tr>
</tbody>
</table>

VIII. ANALYSIS OF EXPERIMENTAL RESULTS

As shown above, contextual polarity recognition task (Table II polar results) enhances the classification of prior polarities of expressions in Table I. As well, the selected features surpasses both baseline classifiers (Table II and Table III). The final output of this research is that SentiRDI augmented with contextual polarities and the related phrases or examples; APPENDIX A shows some samples of our output.

From our experiments, we found that the quality of the prior polarity and the contextual polarity depend on many pre-required Natural Language Processing (NLP) tasks. These tasks are very useful to acquire prior and contextual polarities of the subjective clues, unfortunately, they add as well, at the same time, incremental error ratios to our target mission. The pre-required NLP tasks are:-

**Normalization of writing Arabic:** in Arabic language there are some letters have different forms. For example, ["Alif"] ["ا"] has four forms ["ا", "أ", "۰", "ً"]. ["Yaa"] has two forms ["ي" , "ي" ] and ["Taa el marpouta" and el haa el marpouta"] ["ي", "ي"].

**Arabic parser:** unfortunately, until now there exists no highly accurate public parser for Arabic language due to its high ineffectual nature, complexity, and variant sources of ambiguities (lexical, structural, and semantic).

**Named Entity Recognition:** we used only the named entities extracted by [8] so we were dramatically affected by its performance.
IX. Conclusion and Future Work

In this paper, we study the seminal English two-step contextual polarity phrase-level recognition approach [3] to enhance word polarities within its contexts in Arabic language. Using this approach, we are able to automatically identify the contextual polarities for our Arabic large set of sentiment expressions, achieving results that are significantly better than baselines. Our main contribution is to acquire the sentiment Arabic lexical semantic database (SentiRDI) having the word prior polarities coupled with its contextual polarities and the related phrases (APPENDIX A).

In the future, we are going to extend the database depending on further analysis of exiting opinion mining English corpora. We intend to build our own examples and sentences to enrich the classifier performance with Arabic polar and neutral examples.

REFERENCES

APPENDIX A

<table>
<thead>
<tr>
<th>Word</th>
<th>Prior polarity</th>
<th>Context Polarity</th>
<th>Phrase</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;الحرب&quot;</td>
<td>Negative</td>
<td>Positive</td>
<td>War against terrorism</td>
</tr>
<tr>
<td>&quot;الحرب&quot;</td>
<td>Negative</td>
<td>Negative</td>
<td>War against humanity</td>
</tr>
<tr>
<td>&quot;War&quot;</td>
<td>Positive</td>
<td>Negative</td>
<td>&quot;Mugabe won unfairly&quot;</td>
</tr>
<tr>
<td>&quot;Peace&quot;</td>
<td>Positive</td>
<td>Negative</td>
<td>&quot;Paralysis of the peace process&quot;</td>
</tr>
<tr>
<td>&quot;Agree&quot;</td>
<td>Positive</td>
<td>Negative</td>
<td>&quot;Ambassadors did not agree&quot;</td>
</tr>
<tr>
<td>&quot;comfortable&quot;</td>
<td>Positive</td>
<td>Negative</td>
<td>&quot;Feel uncomfortable&quot;</td>
</tr>
<tr>
<td>&quot;Occupation&quot;</td>
<td>Negative</td>
<td>Positive</td>
<td>&quot;Against the occupation, end the occupation&quot;</td>
</tr>
<tr>
<td>&quot;Security&quot;</td>
<td>Positive</td>
<td>Objective</td>
<td>&quot;Organization for Security and Cooperation&quot;</td>
</tr>
<tr>
<td>&quot;Cooperation&quot;</td>
<td>Positive</td>
<td>Objective</td>
<td>&quot;Organization for Security and Cooperation&quot;</td>
</tr>
<tr>
<td>&quot;Pollution&quot;</td>
<td>Negative</td>
<td>Positive</td>
<td>&quot;Combating Pollution&quot;</td>
</tr>
<tr>
<td>&quot;Reform&quot;</td>
<td>Positive</td>
<td>Objective</td>
<td>&quot;Liberal Reform Party&quot;</td>
</tr>
<tr>
<td>&quot;Football Association&quot;</td>
<td>Positive</td>
<td>Objective</td>
<td>&quot;Football Association&quot;</td>
</tr>
<tr>
<td>&quot;Stability&quot;</td>
<td>Positive</td>
<td>Negative</td>
<td>&quot;Instability in the Muslim world&quot;</td>
</tr>
</tbody>
</table>