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, ["ظللت أعمل بجد و أجتهاد طيلة الأشهر الماضية لكن النتائج كانت سيئه"] ["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, ["جيد","سعادة","رائع"] ["good", "happiness"," wonderful"] have positive prior polarities and ["حزن","بغيض","سئ"] ["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:-

["لم يوافق سفراء منظمة الأمن و التعاون من الدول ال 55 الاعضاء في المنظمة على إرسال مراقبين إضافيين بعد أن رفعت روسيا اعتراضاتها على وجود معززين وقال السفير ستودمان انه من الضروري ان تتقيد الدول بالتزاماتها الدولية بالحفاظ على حقوق الإنسان و الحريات الأساسية في الحرب ضد الارهاب."]

["Ambassadors of the Organization for Security and Cooperation of the 55 member states of the organization did not agree to send additional observers after Russia lifted its objections to the presence of reinforcing Stoudmann. The ambassador said that it is necessary to comply with International obligations of states to preserve human rights and fundamental freedoms in the war against terrorism".]

["يوافق ","منظمة ","الأمن","التعاون "," المنظمة","بالحفاظ "," حقوق "," الحريات"]

["agree", "organization", "security", "cooperation", "organization", "maintain", "rights", "freedom"]

The above words have positive prior polarities, but they are not all being used to express positive sentiments. For example, ["يوافق"] ["agree"] is preceded by a negative tool ["لم "] ["not"] so it has a negative contextual polarity. Also, the words ["منظمة ","الأمن ","التعاون","المنظمة","] ["organization", "security", "cooperation", "organization"] have neutral contextual polarities because they are organization names. But the words ["الحريات ","حقوق "," بالحفاظ"] ["preservation", "rights", "freedom"] have similar prior and contextual polarities. Also these words [تتقيد "," اعتراضتها "," الحرب "," الارهاب " "] ["terrorism", "war", "objection", "comply"] have negative prior polarities but they are not all being used to express negative sentiments. For example, the expression ["الحرب ضد الارهاب "] ["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 ["ليس جيدا"] ["not good"] or it may involve long distance dependency such as ["الحضارة الغربية لا تستطيع تكوين نظام أكثر سعاده"] ["Western civilization can't configure a global system happier"]. Intensifiers influence the force of the term [...["كثير","قليل","بالغة","بعمق","جدا"] ["a lot", "a little", "very", "deeply", "too" ...] . Shifter words precede or follow the polar term and influence its polarity, for example, ["فاز ظلما"] ["Won unfairly"], ["تمنع العقوبة"] ["prevent punishment"]. Connectors

also may influence the contextual polarity; there are some connectors give similar polarities for all connected words ["","و","أو"] ["and", "or"] and some connectors express different polarities [..."بالرغم من", "على النقيض", "لكن", "على العكس"] ["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[1] 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)[2] 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.**

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

Where n is number of words in sentence, Pwi polarity of word i in sentence that is specified before from prior polarity database (SentiRDI) such that

$$p_{wi} = \begin{cases} -1 \; negative \\ 0 \; objective \\ 1 \; positive \end{cases} \quad (2)$$

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

$$Avg\ TF\_ISF_{si} \frac{1}{||Si||} \sum_{t=1}^{||si||} (TF_{t,si} * ISF_t) \quad (3)$$

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

$$TF_{t,s} = \frac{N_{t,s}}{||S||} \quad (4)$$

Where Nt,s is the number of occurrences of term t in sentence S. ||S|| 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.

$$\text{ISF} = \log \frac{S}{S_i} \quad (5)$$

Where S is the number of all sentences in the corpus and Si 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 [" فرنسا لن توافق على هذه الصيغة"] ["France will not agree to this formula"] or it may have long distance dependency of the clue as [" بدون مساعدة الولايات المتحدة عبر صندوق النقد الدولي فان الدولة لن تكون قادرة على تحقيق الاستقرار الاقتصادي والاجتماعي والسياسي "] ["Without the help of the United States through the International Monetary Fund, the state will not be able to achieve economic and social stability and political"]. We consider the Arabic language negation tools namely ["لم","ليس","لن","ما","لما","إن","لا","لات"] transliterated in English as ["lam", "lays","ln","maa","lamaa","en","laa","lat"].

Intensifiers: a word that has little meaning itself but provides force, intensity or emphasis to another word. Intensifier may be before or after the subjective clues. Arabic intensifiers examples are ["كثير","قليل","بالغة","بعمق","جدا"] [" a lot", "a little", "very", "deeply", "too"]. We collected a set of intensifiers found in the used corpus and the others translated from Grammar of English Language [6].

Presupposition items: the words shift the valence of evaluative terms through their presuppositions [10]. These words are collected during exploration of the contextual polarity annotations in our development data. Here we divide it into four categories:-

General shifters: the shifters invert the polarity of subjective clue such as ["ضد","وقف", "عدم", "منع"]["prevention", "not", "stop", "against"].

Positive shifters: the shifters change polarity always to positive such as ["تخطي","مكافحة", "صد", "نفي"] ["deny", "bodice", "combat", "skip"].

Negative shifters: these shifters change polarity always to negative such as ["يحظر","يفتقر", "يحرم", "ينقص"] ["decrease", "deprive", "lack", "prohibit"].

Objective shifters: these shifters help to extract Named Entity (NE) from the text such as [" جماعة","صحيفة", "وكالة", "حزب","ولي العهد"] ["Group"," newspaper", "agency", "party", "the Crown Prince"].

The above contextual polarity influencers are extracted from our corpus and used in our classifiers as features as described below. In order to classify the contextual polarities of the subjective expressions, first we determine whether the clue instances are neutral or polar in their contexts. While neutral clues are words which have non-neutral prior polarities with neutral contextual polarities, polar clues are words which have non-neutral prior polarities with non-neutral contextual polarities. Second, all polar clues that result from the first-step are taken for more classification to determine whether the polar clue instance has positive contextual polarity or negative polar polarity.

B. Baseline (prior polarity classifier)

We created a simple prior polarity classifier (TABLE I) assuming that the contextual polarity of a clue instance equals to the clue's prior polarity. We apply this classifier on all extracted subjective expression (18,678) from translated MPQA corpus. The classifier has accuracy of 48.45% and the following table describes the results of this classifier.

TABLE I. BASELINE CLASSIFIER RESULTS

	Positive expression	**Negative Expression**	**all**
F	67.6	42.6	52.4
P	76.6	35	48.45
R	60.1	54.4	57.1

C. Features of Neutral-polar classification

The neutral-polar classifier is to recognize the neutral clues from the polar ones. 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 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.

POS of SC: it is the part of speech of the subjective clue. We used Stanford Log-Linear Part of Speech Tagger to extract POS.

POS of previous word: it is the POS that presents POS tag of the SC previous word.

POS of next word: it is the POS that presents POS tag of the SC next word.

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.

NER_SC: it is the binary feature to present if the subjective clue is a named entity.

SC_before: it is the binary feature to present if the subjective clue is preceded by another one.

SC_After: it is the binary feature to present if the subjective clue is followed by another one.

Self_intensifier: it is the binary feature to present if subjective clue is one of intensifiers or not.

Intensifier_before_after: it is the binary feature to present if there is intensifier before or after the subjective clue.

Connector: it is the binary feature to present if there is connector ["أو" , "و"] ["and", "or"] between two subjective clues (in this case they have the same polarity) .

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

Obj_shifter: it is the binary feature to present if there is one of objective shifters before a subjective clue.

Self_obj_shifter: it is the binary feature to present if the subjective clue is one of objective shifters or not.

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 connector ["أو" , "و"] ["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 ["بالرغم من", "لكن"] ["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 negative tools. Here, we consider a 4–word window before the subjective clue to deal with longer-distance dependencies.

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

Negative polarity shifter: is the binary feature to present if the subjective clue is preceded by one of the shifters; these shifters 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 shifters; these shifters 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

	WT	WT + PP	All	WT	WT+ PP	All
	Polar			Neutral		
F	78.4	87.6	91.5	72	81.8	84.3
P	77	80	86.7	60.5	69.4	88.2
R	80	96.8	96.8	89	99.8	80.7

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

	WT	WT + PP	All	WT	WT+ PP	All
	Positive			Negative		
F	79.2	81.2	81.3	76.2	81	82.4
P	75.7	80.7	80. 8	80.2	80.7	83.1
R	83	81.8	81.9	72.6	81.4	81.6

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; APPENDEX 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"] [“a”] has four forms ["إ","ٱ","أ","آ"], ["Yaa"] has two forms ["ي" , "ى"] and ["Taa el marpouta” and el haa el marpouta "] ["ه","ة"].

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

[1] T. Wilson, J. Wiebe, and P. Hoffman. 2005. "Recognizing contextual polarity in phrase level sentiment analysis". In Proceedings of ACL.

[2] H. Mobarz, M. Rashwan, and S. AbdelRahman, 2011, "Generating lexical Resources for Opinion Mining in Arabic language automatically," The Eleventh Conference on Language Engineering ESOLEC', Cairo-Egypt, Sept.I.

[3] T. Wilson, J. Wiebe, and P. Hoffmann, 2009. "Recognizing contextual polarity:An exploration of features for phrase-level sentiment analysis," Computational linguistics, vol. 35, no. 3, pp. 399–433.

[4] T. Nasukawa and J. Yi. 2003. "Sentiment analysis: Capturing favorability using natural language processing". In K-CAP 2003.

[5] J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack. 2003. "Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques". In IEEE ICDM-2003.

[6] R. Quirk, S. Greenbaum, G. Leech, and J. Svartvik. 1985. A "Comprehensive Grammar of the English Language". Longman,New York.

[7] J. Wiebe and E. Riloff, 2005. "Creating Subjective and Objective Sentence Classifiers from Unannotated Texts," In Proceeding of CICLing-05, International Conference on Intelligent Text Processing and Computational Linguistics, pages 486–497, Mexico City, Mexico.

[8] AbdelRahman, M. Elarnaoty, M. Magdy and A. Fahmy, 2010. "Integrated Machine Learning Techniques for Arabic Named Entity Recognition, "IJCSI International Journal of Computer Science, pp. 1694-0784.

[9] M. Attia, and M. Rashwan, 2004. "A Large Scale Arabic POS Tagger Based on a Compact Arabic POS Tag Set and Application on the Statistical Inference of Syntactic Diacritics of Arabic Text Words ," NEMLAR.

[10] L. Polanya and A. Zaenen. 2004."Contextual valence shifters." In Working Notes of the AAAI Spring Symposium on Exploring Attitude and Affect in Text:Theories and Applications, pages 106–111,

Appendix A

Word	Prior polarity	Context Polarity	Phrase
"الحرب" *"War"*	Negative	Positive	"الحرب ضد الأرهاب" *"War against terrorism "*
"الحرب" *"War"*	Negative	Negative	"الحرب ضد الإنسانية" *"War against humanity "*
"فاز " *"Won"*	Positive	Negative	"موجابى فاز ظلما" *"Mugabe won unfairly "*
"السلام " *"Peace"*	Positive	Negative	"شلل عملية السلام" *"Paralysis of the peace process"*
"يوافق" *"Agree"*	Positive	Negative	"لم يوافق سفراء" *"Ambassadors did not agree "*
"الارتياح" *"comfortable"*	Positive	Negative	"يشعرون بعدم الارتياح" *"Feel uncomfortable "*
"الاحتلال " *"occupation "*	Negative	Positive	"انهاء الاحتلال , ضد الاحتلال" *"Against the occupation, end the occupation "*
"الأمن " *"Security "*	Positive	Objective	"منظمة الأمن والتعاون" *" Organization for Security and Cooperation "*
"التعاون " *"Cooperation "*	Positive	Objective	منظمة الأمن والتعاون *" Organization for Security and Cooperation "*
Pollution" التلوث	Negative	Positive	"مكافحة التلوث" *"Combating Pollution "*
"الاصلاح " *"Reform"*	Positive	Objective	"حزب الاصلاح الليبرالي" *"Liberal Reform Party"*
"كره" "hate"	Negative	Objective	"اتحاد كره القدم" *"Football Association "*
"الاستقرار " *"stability"*	Positive	Negative	"تزعزع الاستقرار في العالم الإسلامي" *" Instability in the Muslim world "*