Generating lexical Resources for Opinion Mining in Arabic language automatically

Hanaa Bayomi Ali [*1], Mohsen Rashwan [**2], Samir Abd_Elrahman [*3]

**computer Science Department, Faculty of Computers and Information and Cairo University*
Giza 12613, Egypt
[1]h.mobarz@fci-cu.edu.eg
[3]s.abdelrahman@fci-cu.edu.eg

[**] *Electronics and Communications Department, Faculty of Engineering and Cairo University*
The Engineering Company for the Development of Computer Systems; RDI, Al-Haram Av., 12111, Giza, Egypt
[2] Mohsen_Rashwan@RDI-eg.com

***Abstract*— In this work we present SENTIRDI, a lexical resource explicitly devised for supporting sentiment classification and opinion mining applications. We confront the task of deciding whether a given Arabic term has a positive connotation, or a negative connotation, or has no subjective connotation at all; this problem thus subsumes the problem of determining subjectivity and the problem of determining orientation. We tackle this problem by bootstrapping from three small sets of terms (Positive, Objective, and Negative seed sets) and increase sets consequently by applying lexical relation that is available in RDI Lexical Semantic Data Base(RDILSDB) until cover all Arabic semantic fields.**

1 INTRODUCTION

Opinion mining is a recent sub discipline of computational linguistics which is concerned not with the topic a document is about, but with the opinion it expresses. Opinion-driven content management has several important applications, such as determining critics' opinions about a given product by classifying online product reviews, or tracking the shifting attitudes of the general public toward a political candidate by mining online forums.

Within opinion mining process several tasks are defined these tasks involve tagging a given document depending on the opinion it express. The defined tasks are:-

- **Determining document subjectivity**, as in deciding whether a given text has a factual nature (i.e. describes a given situation or event, without expressing a positive or a negative opinion on it) or expresses an opinion on its subject matter. This amounts to performing binary text categorization under categories Objective and Subjective (Pang and Lee,2004; Yu and Hatzivassiloglou, 2003); ([1]; [2]);

- **Determining document orientation (or polarity)**, as in deciding if a given Subjective text expresses a positive or a negative opinion on its subject matter (Pang and Lee, 2004;Turney, 2002); ([1]; [3]);

- **Determining the strength of document orientation**, as in deciding whether the positive opinion expressed by a text is weakly positive, mildly positive, or strongly positive (Wilson et al., 2004) ([4]).

In order to aid these tasks, we need to identify the orientation of subjective terms contained in text, i.e. determining whether a term that carries opinionated content has a positive or a negative connotation.

Opinion Mining for Arabic language is considered a hot research topic with a very few contributions. This is due to the complexity of Arabic language and rareness of Opinion Mining Arabic Linguistic resources. This problem thus subsumes the problem of determining subjectivity and the problem of determining orientation.

One of the big challenges while dealing with term orientation in Arabic language for which one would like to perform opinion mining is that, there is no available lexical resource where terms are tagged as having either positive or negative connotation. the absence of such a resource emerged the need to generate it automatically.

2 RELATED WORK

A. Determining term orientation

Most previous work dealing with the properties of terms within an opinion mining perspective focused on determining term orientation.

Hatzivassiloglou and McKeown (1997) [5] attempt to predict the orientation of subjective adjectives by analysing pairs of adjectives (conjoined by 'and', 'or', 'but', 'either-or', or 'neither-nor') extracted from a large unlabelled document set. The underlying intuition is that the act of conjoining adjectives is subject to linguistic constraints on the orientation of the adjectives involved; e.g. 'and' usually conjoins adjectives of equal orientation, while 'but' conjoins adjectives of opposite orientation. The authors generate a graph where terms are nodes connected by "equal-orientation" or "opposite-orientation" edges, depending on the conjunctions extracted from the document set. A clustering algorithm then partitions the graph into a Positive cluster and a Negative cluster, based on a relation of similarity.

Turney and Littman (2003) [6] determine term orientation by bootstrapping from two small sets of subjective "seed" terms (with the seed set for Positive containing terms such as good and nice, and the seed set for Negative containing terms such as bad and nasty). Their method is based on computing the point wise mutual information (PMI) of the target term t with each seed term ti as a measure of their semantic association. Given a target term t, its orientation value O(t) (where positive value means positive Orientation, and higher absolute value means stronger orientation) is given by the sum of the weights of its semantic association with the seed positive terms minus the sum of the weights of its semantic association with the seed negative terms. For computing PMI, term frequencies and co-occurrence frequencies are measured by querying a document set by means of the AltaVista search engine with a "t" query, a "ti " query, and a "t NEAR ti" query, and using the number of matching documents returned by the search engine as estimates of the probabilities needed for the computation of PMI.

Kamps et al. (2004) [7] consider instead the graph defined on adjectives by the WordNet 2 synonymy relation, and determine the orientation of a target adjective t contained in the graph by comparing the lengths of (i) the shortest path between t and the seed term good, and (ii) the shortest path between t and the seed term bad: if the former is shorter than the latter, than t is deemed to be Positive, otherwise it is deemed to be Negative.

Takamura et al. (2005) [8] determines term orientation (for Japanese) according to a "spin model",i.e. a physical model of a set of electrons each endowed with one between two possible spin directions, and where electrons propagate their spin direction to neighbouring electrons until the system reaches a stable configuration. The authors equate terms with electrons and term orientation to spin direction. They build a neighbourhood matrix connecting each pair of terms if one appears in the gloss of the other, and iteratively apply the spin model on the matrix until a "minimum energy" configuration is reached. The orientation assigned to a term then corresponds to the spin direction assigned to electrons

The system of Kim and Hovy (2004) [9] tackled orientation detection by attributing, to each term, a positivity score and a negativity score; interestingly, terms may thus be deemed to have both positive and negative correlation, maybe with different degrees, and some terms may be deemed to carry a stronger positive (or negative) orientation than others. Their system starts from a set of positive and negative seed terms, and expands the positive (resp. negative) seed set by adding to it the synonyms of positive (resp. negative) seed terms and the antonyms of negative (resp. positive) seed terms. The system classifies then a target term t into either positive or negative by means of two alternative learning-free methods based on the probabilities that synonyms of t also appear in the respective expanded seed sets. A problem with this method is that it can classify only terms that share some synonyms with the expanded seed sets.

The method of (Esuli and Sebastiani, 2005) [10] starts from two small seed (i.e. training) sets Lp and Ln of known positive and negative terms, respectively, and expands them into the two final training sets Trp $\supset$ Lp and Trn $\supset$ Ln by adding them new sets of terms up and unfound by navigating the WordNet graph along the synonymy and antonymy relations. This process is based on the hypothesis that synonymy and antonymy, in addition to defining a relation of meaning, also define a relation of orientation, i.e. that two synonyms typically have the same orientation and two antonyms typically have opposite orientation.

When tested on the same benchmarks, the methods of (Esuli and Sebastiani, 2005; Turney and Littman, 2003) [10,6] performed with comparable accuracies (however, the method of (Esuli and Sebastiani, 2005) [10] is much more efficient than the one proposed by (Turney and Littman, 2003) [6]) , and have outperformed the method of (Hatzivas-siloglou and McKeown, 1997) [5] by a wide margin and the one by (Kamps et al., 2004) [7] by a very wide margin. The methods described in (Hatzi-vassiloglou and McKeown, 1997) [5] is also limited by the fact that it can only decide the orientation of adjectives, while the method of (Kamps et al., 2004) [7] is further limited in that it can only work on adjectives that are present in WorldNet. The methods of (Kim and Hovy, 2004; Takamura et al., 2005) [9,8] are difficult to be compared with the other methods since they were not evaluated on publicly available datasets.

B. Determining term subjectivity

Riloff et al. (2003) [11] developed a method to determine whether a term has a subjective or an objective connotation, based on bootstrapping algorithms. The method identifies patterns for the extraction of subjective nouns from text, bootstrapping from a seed set of 20 terms that the authors judge to be strongly subjective and have found to have high frequency in the text collection from which the subjective nouns must be extracted. The results of this method are not easy to compare with the ones we present in this paper because of the different evaluation methodologies. While we adopt the evaluation methodology used in all of the papers reviewed so far (i.e. checking how good our system is at replicating an existing, independently motivated lexical resource), the authors do not test their method on an independently identified set of labelled terms, but on the set of terms that the algorithm itself extracts. This evaluation methodology only allows testing precision, and not accuracy tout court, since no quantification can be made of false negatives (i.e. the subjective terms that the algorithm should have spotted but has not spotted). This will prevent us from drawing comparisons between this method and our own.

Baroni and Vegnaduzzo (2004) [12] apply the PMI method first used by Turney and Littman (2003) [6] to determine term orientation and subjectivity. Their method uses a small set Ss of 35 adjectives, marked as subjective by human judges, to assign a subjectivity score to each adjective to be classified. Therefore, their method, unlike our own, does not classify terms (i.e. take firm classification decisions), but ranks them according to a subjectivity score, on which they evaluate precision at various level of recall.

C. Multilingual Sentiment Analysis

There is a growing body of work on multilingual sentiment analysis. Most approaches focus on resource adaptation from one language (usually English) to another with few sentiment resources. Mihalcea et al. (2007)[13], for example, generate subjectivity analysis resources in a new language from the English sentiment resources by leveraging a bilingual dictionary or a parallel corpus. Banea et al. (2008; 2010) [14,15]instead automatically translate the English resources by using automatic machine translation engines for subjectivity classification. Prettenhofer and Stein (2010)[16] investigate cross-lingual sentiment classification from the perspective of domain adaptation based on structural correspondence learning (Blitzer et al., 2006)[17].
Approaches that do not explicitly involve resource adaptation include (Wan (2009))[18], which uses co-training (Blum and Mitchell, 1998)[19] with English vs. Chinese features comprising the two independent "views" to exploit unlabeled Chinese data and a labeled English corpus and thereby improves Chinese sentiment classification.

Another notable approach is the work of Boyd-Graber and Resnik (2010)[20], in which they present a generative model, supervised multilingual latent Dirichlet allocation, by jointly modeling topics that are consistent across languages, and employing them to better predict sentiment ratings.
Unlike the methods described above, we focus on simultaneously improving the performance of sentiment classification in a pair of languages by developing a model that relies on sentiment-labelled data in each language as well as unlabelled parallel text for the language pair.

D. SentiWordNet

SentiWordNet, a lexical resource produced by asking an automated classifier $\hat{\Phi}$ to associate to the unique sense represented by each synset s of WordNet (version 2.0) a triplet of scores $\hat{\Phi}(s, p)$ (for $p \in P = \{\text{Positive, Negative, Objective}\}$) describing how strongly that sense enjoy each of the three properties. The method used to develop SentiWordNet is based on the term classification . The score triplet is derived by combining the results produced by a committee of eight ternary classifiers, all characterized by similar accuracy levels but extremely different classification behaviours. (Esuli and Sebastiani, 2005)[10]

3 DETERMINING SUBJECTIVITY & ORIENTATION OF ARABIC TERMS

We present a method for determining term orientation and term subjectivity using semi-supervised technique. Our process is composed of the following steps:-

1. Seed sets (S_p, S_n, S_o) represent three seed sets one for a positive set, one for a negative set and the last for objective set. They are provided as input.

2. Apply lexical Relation (Causality (القحط و الإهلاك), (Causative (شدة الحُب و الألفة), antonym (الحب و الكره), hyponymy (الخيانة و الذنب) and Hypernym (e. g. الجريمة و الخيانة)) from a Lexical semantic Data Base for

   each term in (S_p, S_n) ,in order to find new semantic fields and increase the size of seed set . The new semantic fields, once added to the original ones, yield two new, richer sets S_p and S_n of semantic fields.
3. The new semantic fields that produce from antonym relation have an opposite orientation otherwise have the same orientation.
4. Iterate step 2 and 3 until there are no other new semantic fields that can be added in S_p or S_n.
5. For objective seed set (So) apply the following lexical relation Totality (e.g. الرجل و القدم), Part_of (e.g. القدم و الرجل), Inclusion_K (e.g. الألة و المنجنيق), KindOf (e.g. الألة و المنجنيق), Circumstantial_Place (e.g. و أرض المحشر الإنسان) , Locality_Place (e.g. الأرض الطينية وآلة رش الماء), Circumstantial_Time (e.g. الصوت و الزمن المحدد) and Locality_Time (e.g. الأمس و النوم) for each term in (S_o) in order to find new semantic fields. The new semantic fields, once added to the original ones, yield new set S_o.
6. Iterate step 5 until there is no other new semantic fields that can be added in S_o.
7. We classify common terms among three produced sets (S_p, S_n, S_o) under some criteria (e.g. Shortest Path, Number Of Repetition) .in order to eliminate redundancy among three sets.

Finally we obtain three sets of Semantic fields each one contain all semantic fields with the same polarity.

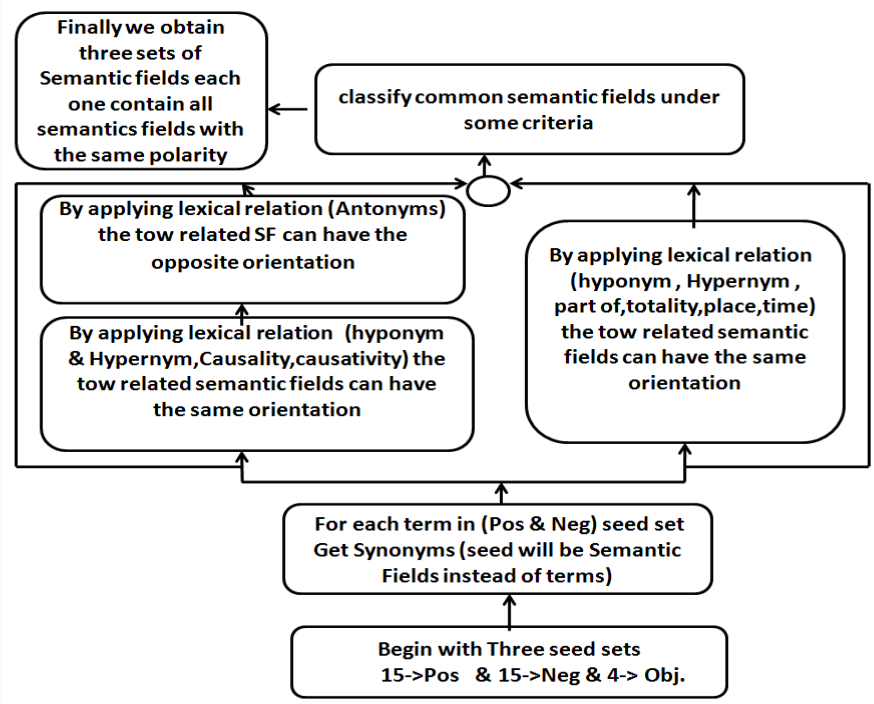

Figure1: represent proposed algorithm for determining term orientation and term subjectivity.

4 EXPERIMENTS AND TRIALS

A. Seed Sets

We have three seed sets :-

1. Positive seed set Sp which have 15 terms ["∀الألفة, ∀التواضع, ∀الشجاعة, ∀الشهامة, ∀الكرَم, ∀الجودة, ∀العطاء, ∀الضحك, ∀الجمال, ∀الصدق, ∀السهل, ∀الذكاء, ∀الاجتهاد, ∀التهذيب, ∀الأمل"] .
2. Negative seed set Sn which have 15 terms ["∀الخوف, ∀الفقر, ∀التمرد, ∀التشاؤم, ∀الضعف, ∀الفاسد, ∀الحزن, ∀الكره, ∀البخل, ∀الإبكاء, ∀التسول, ∀الذنب, ∀القبح, ∀الجهل, ∀الكدر"].
3. Objective seed set So which have 4 terms [""∀الآلة , ∀الأرض, ∀الرجل, ∀الصوت"].

The method requires bootstrapping from a seed sets (Sp, Sn, So) representative of the categories Positive, Negative and Objective. In our experiments we begin from Turney and Littman (2003), (Esuli and Sebastiani, 2005) ([10],[6])seed sets. In order to classify the largest number of semantic fields in Semantic Data Base we increase the number of terms gradually in seed sets by noticing the results until we reach the satisfied result.

For objective seed set, we should notice that in previous work objective terms can be concluded from positive and negative terms [10] but here we begin from seed set in order to improve the results.

B. Expansion method for seed sets

We use RDI Lexical Semantic Data Base (RDILSDB) as the source of lexical relations. (RDILSDB) contains 18.413 semantic fields. It covers 100.000 words. Semantic fields relate together with 293,000 Bilateral semantic via 20 lexical relations (Part_of, Totality, Inclusion_K, KindOf, Inclusion_M, Member_of, Inclusion_I, Integerartion, Inclusion_O, Original, Conditional, Required_Condition, Causality, Causative, Circumstantial_Place, Locality_Place, Circumstantial_Time, Locality_Time, synonym and Antonomy).

5 RESULTS

In fact, fully testing the accuracy of our tagging method experimentally is impossible, since this would require a version of all semantic fields in Arabic language manually annotated according to our three properties of interest, and the unavailability of such a manually annotated resource is exactly the reason why we are interested in generating it automatically.
Our proposed work is evaluated by two methods:-
The first method is using a manually annotated subset of Arabic semantic fields as a “gold standard”. it is annotated by 5 annotator (3 Linguistics from an engineering company to develop digital system (RDI) and 2 Linguistics from Faculty of Dar Science, Cairo university)
Number of Semantic Fields in a “gold standard” is 7216. Table1 represent recall and precision of positive, negative, objective and all semantic fields' regardless polarity after applying proposed algorithm.
The second method is Translating The Micro-WNOp[S. Cerini, V. Compagnoni, A. Demontis, M 2007] (by Google translator) gold standard that is used to evaluate the Senti wordnet which contains 1.105 synset. Table 2 represents recall and precision of positive, negative, objective and all semantic fields' regardless polarity after applying proposed algorithm.

TABLE1

REPRESENT RECALL AND PRECISION OF POSITIVE, NEGATIVE, OBJECTIVE AND ALL SEMANTIC FIELDS FOR THE FIRST TEST.

	Pos_SF	**Neg_SF**	**Obj_SF**	**All_SF**
Recall	86.44	83.16	89.3	87.32
Precision	74.76	79.95	93.09	87.72
F-measure	80.18	81.52	91.16	87.52

TABLE2

REPRESENT RECALL AND PRECISION OF POSITIVE, NEGATIVE, OBJECTIVE AND ALL SEMANTIC FIELDS FOR THE SECOND TEST.

	Pos_SF	**Neg_SF**	**Obj_SF**	**All_SF**
Recall	79,43	85	89	84.67
Precision	80	82,45	87.93	84.95
F-measure	79,71	83.7	88.46	84.81

6 CONCLUSIONS

We have presented a method for determining both term subjectivity and term orientation for using semi-supervised technique which can be considered a very useful tool for all Arabic opinion mining applications because of its wide coverage (all LSDB Semantic fields are tagged according to each of the three labels Objective, Positive and Negative).

It is found that the direct translation from one language (English) to other language (Arabic) doesn't give accurate results because the same term may have a lot of meaning with different polarity.(Ex. Ball has the following meaning كرة, حفلة راقصة ,نزهة , رصاصة ,جسم مستدير من الإنسان ,لعبة من ألعاب الكرة)

After a lot of experiments it is found that (RDILSDB) doesn't recognize on countries nouns, numbers and Currency On the contrary Wordnet does.

ACKNOWLEDGMENT

This work was partially supported by Project ICT Centers of Excellence (CoE) “Data Mining and Computer Modeling”, funded by Itida.

REFERENCES

[1]	*Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of ACL-04, 42nd Meeting of the Association for Computational Linguistics,pages 271–278, Barcelona, ES.*
[2]	*Hong Yu and Vasileios Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In Proceedings of EMNLP-03, 8th Conference on Empirical Methods in Natural Language Processing, pages 129–136, Sapporo, JP.*
[3]	*Peter Turney. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In Proceedings of ACL-02, 40th Annual Meeting of the Association for Computational Linguistics, pages 417–424, Philadelphia, US.*
[4]	*Theresa Wilson, Janyce Wiebe, and Rebecca Hwa. 2004.Just how mad are you? Finding strong and weak opinion clauses. In Proceedings of AAAI-04, 21st Conference of the American Association for Artificial Intelligence, pages 761–769, San Jose, US.*
[5]	*Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1997.Predicting the semantic orientation of adjectives. In Proceedings of ACL-97, 35th Annual Meeting of the Association for Computational Linguistics, pages 174–181,Madrid, ES.*
[6]	*Peter D. Turney and Michael L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems, 21(4):315–346.*
[7]	*Jaap Kamps, Maarten Marx, Robert J. Mokken, and Maarten De Rijke. 2004. Using WordNet to measure semantic orientation of adjectives. In Proceedings of LREC-04, 4th International Conference on Language Resources and Evaluation, volume IV, pages 1115–1118, Lisbon, PT.*
[8]	*Hiroya Takamura, Takashi Inui, and Manabu Okumura 2005. Extracting emotional polarity of words using spin model. In Proceedings of ACL-05, 43rd Annual Meeting of the Association for Computational Linguistics, pages 133–140, Ann Arbor, US.*
[9]	*Soo-Min Kim and Eduard Hovy. 2004. Determining the sentiment of opinions. In Proceedings of COLING-04, 20th International Conference on Computational Linguistics,pages 1367–1373, Geneva, CH.*

[10]	*Andrea Esuli and Fabrizio Sebastiani. 2005. Determining the semantic orientation of terms through gloss analysis. In Proceedings of CIKM-05, 14th ACM International Conference on Information and Knowledge Management, pages 617–624, Bremen, DE.*
[11]	*Ellen Riloff, Janyce Wiebe, and Theresa Wilson. 2003.Learning subjective nouns using extraction pattern bootstrapping. In Proceedings of CONLL-03, 7th Conference on Natural Language Learning, pages 25–32, Edmonton, CA.*
[12]	*M. Baroni and S. Vegnaduzzo. 2004. Identifying subjective adjectives through Web-based mutual information. InProceedings of KONVENS-04, 7th Konferenz zur Verarbeitung* *Nat¨urlicher Sprache (German Conference on Natural Language Processing), pages 17–24, Vienna, AU.*
[13]	*Rada Mihalcea, Carmen Banea, and Janyce Wiebe. 2007. Learning multilingual subjective language via cross-lingual projections. In Proceedings of ACL'07.*
[14]	*Carmen Banea, Rada Mihalcea, and Janyce Wiebe. 2010. Multilingual subjectivity: Are more languages better? In Proceedings of COLING'10.*
[15]	*Carmen Banea, Rada Mihalcea, Janyce Wiebe, and Samer Hassan. 2008. Multilingual subjectivity analysis using machine translation. In Proceedings of EMNLP'08.*
[16[	*Peter Prettenhofer and Benno Stein. 2010. Cross-language text classification using structural correspondence learning. In Proceedings of ACL'10.*
[17]	*John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspond-dence learning. In Proceedings of EMNLP'06.*
[18]	*Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. In Proceedings of ACL/AFNLP'09.*
[19]	*Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In Proceedings of COLT'98.*
[20]	*Jordan Boyd-Graber and Philip Resnik. 2010. Holistic sentiment analysis across languages: Multilingual supervised Latent Dirichlet Allocation. In Proceedings of EMNLP'10.*