Subjectivity and Sentiment Analysis of Modern Standard Arabic

Muhammad Abdul-Mageed
Department of Linguistics &
School of Library & Info. Science,
Indiana University,
Bloomington, USA,
mabdulma@indiana.edu

Mona T. Diab
Center for Computational
Learning Systems,
Columbia University, NYC, USA,
mdiab@ccls.columbia.edu

Mohammed Korayem
School of Informatics
and Computing,
Indiana University,
Bloomington, USA,
mkorayem@indiana.edu

Abstract

Although *Subjectivity and Sentiment Analysis (SSA)* has been witnessing a flurry of novel research, there are few attempts to build SSA systems for Morphologically-Rich Languages (MRL). In the current study, we report efforts to partially fill this gap. We present a newly developed manually annotated corpus of Modern Standard Arabic (MSA) together with a new polarity lexicon.The corpus is a collection of newswire documents annotated on the sentence level. We also describe an automatic SSA tagging system that exploits the annotated data. We investigate the impact of different levels of preprocessing settings on the SSA classification task. We show that by explicitly accounting for the rich morphology the system is able to achieve significantly higher levels of performance.

1 Introduction

Subjectivity and Sentiment Analysis (SSA) is an area that has been witnessing a flurry of novel research. In natural language, *subjectivity* refers to expression of opinions, evaluations, feelings, and speculations (Banfield, 1982; Wiebe, 1994) and thus incorporates *sentiment*. The process of *subjectivity classification* refers to the task of classifying texts into either *objective* (e.g., *Mubarak stepped down*) or *subjective* (e.g., *Mubarak, the hateful dictator, stepped down*). Subjective text is further classified with *sentiment* or *polarity*. For sentiment classification, the task refers to identifying whether the subjective text is *positive* (e.g., *What an excellent camera!*), *negative* (e.g., *I hate this camera!*), *neutral* (e.g., *I believe there will be a meeting.*), or, sometimes, *mixed* (e.g., *It is good, but I hate it!*) texts.

Most of the SSA literature has focused on English and other Indio-European languages. Very few studies have addressed the problem for morphologically rich languages (MRL) such as Arabic, Hebrew, Turkish, Czech, etc. (Tsarfaty et al., 2010). MRL pose significant challenges to NLP systems in general, and the SSA task is expected to be no exception. The problem is even more pronounced in some MRL due to the lack in annotated resources for SSA such as labeled corpora, and polarity lexica.

In the current paper, we investigate the task of sentence-level SSA on *Modern Standard Arabic (MSA)* texts from the newswire genre. We run experiments on three different pre-processing settings based on tokenized text from the Penn Arabic Treebank (PATB) (Maamouri et al., 2004) and employ both language-independent and Arabic-specific, morphology-based features. Our work shows that explicitly using morphology-based features in our models improves the system's performance. We also measure the impact of using a wide coverage polarity lexicon and show that using a tailored resource results in significant improvement in classification performance.

2 Approach

To our knowledge, no SSA annotated MSA data exists. Hence we decided to create our own SSA annotated data.[1]

2.1 Data set and Annotation

Corpus: Two college-educated native speakers of Arabic annotated 2855 sentences from Part 1 V 3.0 of the PATB. The sentences make up the first 400 documents of that part of PATB amounting to a total of 54.5% of the PATB Part 1 data set. For each sentence, the annotators assigned one of 4 possible labels: (1) OBJECTIVE (OBJ), (2) SUBJECTIVE-POSITIVE (S-POS), (3) SUBJECTIVE-NEGATIVE (S-NEG), and (4) SUBJECTIVE-NEUTRAL (S-NEUT). Following (Wiebe et al., 1999), if the primary goal

[1]The data may be obtained by contacting the first author.

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, pages 587–591,
Portland, Oregon, June 19-24, 2011.

of a sentence is judged as the objective reporting of information, it was labeled as OBJ. Otherwise, a sentence would be a candidate for one of the three SUBJ classes. Inter-annotator agreement reached 88.06%.[2] The distribution of classes in our data set was as follows: 1281 OBJ, a total of 1574 SUBJ, where 491 were deemed S-POS, 689 S-NEG, and 394 S-NEUT. Moreover, each of the sentences in our data set is manually labeled by a domain label. The domain labels are from the newswire genre and are adopted from (Abdul-Mageed, 2008).

Polarity Lexicon: We manually created a lexicon of 3982 adjectives labeled with one of the following tags {*positive, negative, neutral*}. The adjectives pertain to the newswire domain.

2.2 Automatic Classification

Tokenization scheme and settings: We run experiments on gold-tokenized text from PATB. We adopt the PATB+Al tokenization scheme, where proclitics and enclitics as well as Al are segmented out from the stem words. We experiment with three different pre-processing lemmatization configurations that specifically target the stem words: (1) *Surface*, where the stem words are left as is with no further processing of the morpho-tactics that result from the segmentation of clitics; (2) *Lemma*, where the stem words are reduced to their lemma citation forms, for instance in case of verbs it is the 3rd person masculine singular perfective form; and (3) *Stem*, which is the surface form minus inflectional morphemes, it should be noted that this configuration may result in non proper Arabic words (a la IR stemming). Table 1 illustrates examples of the three configuration schemes, with each underlined.

Features: The features we employed are of two main types: Language-independent features and Morphological features.

Language-Independent Features: This group of features has been employed in various SSA studies.

Domain: Following (Wilson et al., 2009), we apply a feature indicating the *domain* of the document to which a sentence belongs. As mentioned earlier, each sentence has a document domain label manually associated with it.

[2]A detailed account of issues related to the annotation task will appear in a separate publication.

UNIQUE: Following Wiebe et al. (2004) we apply a *unique* feature. Namely words that occur in our corpus with an absolute frequency < 5, are replaced with the token "UNIQUE".

N-GRAM: We run experiments with *N*-grams ≤ 4 and all possible combinations of them.

ADJ: For subjectivity classification, we follow Bruce & Wiebe's (1999) in adding a binary *has_adjective* feature indicating whether or not any of the adjectives in our manually created polarity lexicon exists in a sentence. For sentiment classification, we apply two features, *has_POS_adjective* and *has_NEG_adjective*, each of these binary features indicate whether a POS or NEG adjective occurs in a sentence.

MSA-Morphological Features: MSA exhibits a very rich morphological system that is templatic, and agglutinative and it is based on both derivational and inflectional features. We explicitly model morphological features of *person*, *state*, *gender*, *tense*, *aspect*, and *number*. We do not use POS information. We assume undiacritized text in our models.

2.3 Method: Two-stage Classification Process

In the current study, we adopt a two-stage classification approach. In the first stage (i.e., *Subjectivity*), we build a binary classifier to sort out OBJ from SUBJ cases. For the second stage (i.e., *Sentiment*) we apply binary classification that distinguishes S-POS from S-NEG cases. We disregard the neutral class of S-NEUT for this round of experimentation. We use an SVM classifier, the SVM^{light} package (Joachims, 2008). We experimented with various kernels and parameter settings and found that linear kernels yield the best performance. We ran experiments with *presence* vectors: In each sentence vector, the value of each dimension is binary either a 1 (regardless of how many times a feature occurs) or 0.

Experimental Conditions: We first run experiments using each of the three lemmatization settings *Surface, Lemma, Stem* using various *N*-grams and *N*-gram combinations and then iteratively add other features. The morphological features (i.e., *Morph*) are added only to the *Stem* setting. Language-independent features (i.e., from the following set {*DOMAIN, ADJ, UNIQUE*}) are added to the *Lemma* and *Stem+Morph* settings. With all

Word	POS	Surface form	Lemma	Stem	Gloss
AlwlAyAt	Noun	Al+wlAyAt	Al+wlAyp	Al+wlAy	the states
ltblgh	Verb	l+tblg+h	l+>blg+h	l+blg+h	to inform him

Table 1: Examples of word lemmatization settings

the three settings, clitics that are split off words are kept as separate features in the sentence vectors.

3 Results and Evaluation

We divide our data into 80% for 5-fold cross-validation and 20% for test. For experiments on the test data, the 80% are used as training data. We have two settings, a development setting (DEV) and a test setting (TEST). In the development setting, we run the typical 5 fold cross validation where we train on 4 folds and test on the 5th and then average the results. In the test setting, we only ran with the best configurations yielded from the DEV conditions. In TEST mode, we still train with 4 folds but we test on the test data exclusively, averaging across the different training rounds.

It is worth noting that the test data is larger than any given dev data (20% of the overall data set for test, vs. 16% for any DEV fold). We report results using *F*-measure (*F*). Moreover, for TEST we report only experiments on the *Stem+Morph* setting and *Stem+Morph+ADJ*, *Stem+Morph+DOMAIN*, and *Stem+Morph+UNIQUE*. Below, we only report the best-performing results across the *N*-GRAM features and their combinations. In each case, our baseline is the majority class in the training set.

3.1 Subjectivity

Among all the lemmatization settings, the *Stem* was found to perform best with 73.17% *F* (with 1g+2g), compared to 71.97% *F* (with 1g+2g+3g) for *Surface* and 72.74% *F* (with 1g+2g) for *Lemma*. In addition, adding the inflectional morphology features improves classification (and hence the *Stem+Morph* setting, when ran under the same 1g+2g condition as the *Stem*, is better by 0.15% *F* than the *Stem* condition alone). As for the language-independent features, we found that whereas the *ADJ* feature does not help neither the *Lemma* nor *Stem+Morph* setting, the *DOMAIN* feature improves the results slightly with the two settings. In addition, the *UNIQUE* feature helps classification with the *Lemma*, but it hurts with the *Stem+Morph*.

Table 2 shows that although performance on the test set drops with all settings on *Stem+Morph*, results are still at least 10% higher than the bseline. With the *Stem+Morph* setting, the best performance on the TEST set is 71.54% *F*and is 16.44% higher than the baseline.

3.2 Sentiment

Similar to the subjectivity results, the *Stem* setting performs better than the other two lemmatization scheme settings, with 56.87% *F* compared to 52.53% *F* for the *Surface* and 55.01% *F* for the *Lemma*. These best results for the three lemmatization schemes are all acquired with 1g. Again, adding the morphology-based features helps improve the classification: The *Stem+Morph* outperforms *Stem* by about 1.00% *F*. We also found that whereas adding the *DOMAIN* feature to both the *Lemma* and the *Stem+Morph* settings improves the classification slightly, the *UNIQUE* feature only improves classification with the *Stem+Morph*.

Adding the *ADJ* feature improves performance significantly: An improvement of 20.88% *F* for the *Lemma* setting and 33.09% *F* for the *Stem+Morph* is achieved. As Table 3 shows, performance on test data drops with applying all features except *ADJ*, the latter helping improve performance by 4.60% *F*. The best results we thus acquire on the 80% training data with 5-fold cross validation is 90.93% *F* with 1g, and the best performance of the system on the test data is 95.52% *F* also with 1g.

4 Related Work

Several sentence- and phrase-level SSA systems have been built, e.g., (Yi et al. 2003; Hu and Liu., 2004; Kim and Hovy., 2004; Mullen and Collier 2004; Pang and Lee 2004; Wilson et al. 2005; Yu and Hatzivassiloglou, 2003). Yi et al. (2003) present an NLP-based system that detects all ref-

	Stem+Morph	+ADJ	+DOMAIN	+UNIQUE
DEV	**73.32**	**73.30**	**73.43**	**72.92**
TEST	**65.60**	**71.54**	**64.67**	**65.66**
Baseline	55.13	55.13	55.13	55.13

Table 2: Subjectivity results on Stem+Morph+language independent features

	Stem+Morph	+ADJ	+DOMAIN	+UNIQUE
DEV	**57.84**	**90.93**	**58.03**	**58.22**
TEST	**52.12**	**95.52**	**53.21**	**51.92**
Baseline	58.38	58.38	58.38	58.38

Table 3: Sentiment results on Stem+Morph+language independent features

erences to a given subject, and determines sentiment in each of the references. Similar to (2003), Kim & Hovy (2004) present a sentence-level system that, given a topic detects sentiment towards it. Our approach differs from both (2003) and Kim & Hovy (2004) in that we do not detect sentiment toward specific topics. Also, we make use of *N*-gram features beyond unigrams and employ elaborate *N*-gram combinations.

Yu & Hatzivassiloglou (2003) build a document- and sentence-level subjectivity classification system using various *N*-gram-based features and a polarity lexicon. They report about 97% F-measure on documents and about 91% F-measure on sentences from the *Wall Street Journal* (WSJ) corpus. Some of our features are similar to those used by Yu & Hatzivassiloglou, but we exploit additional features. Wiebe et al. (1999) train a sentence-level probabilistic classifier on data from the WSJ to identify subjectivity in these sentences. They use POS features, lexical features, and a paragraph feature and obtain an average accuracy on subjectivity tagging of 72.17%. Again, our feature set is richer than Wiebe et al. (1999).

The only work on Arabic SSA we are aware of is that of Abbasi et al. (2008). They use an entropy weighted genetic algorithm for both English and Arabic Web forums at the document level. They exploit both syntactic and stylistic features. Abbasi et al. use a root extraction algorithm and do not use morphological features. They report 93.6% accuracy. Their system is not directly comparable to ours due to the difference in data sets and tagging granularity.

5 Conclusion

In this paper, we build a sentence-level SSA system for MSA contrasting language independent only features vs. combining language independent and language-specific feature sets, namely morphological features specific to Arabic. We also investigate the level of stemming required for the task. We show that the *Stem* lemmatization setting outperforms both *Surface* and *Lemma* settings for the SSA task. We illustrate empirically that adding language specific features for MRL yields improved performance. Similar to previous studies of SSA for other languages, we show that exploiting a polarity lexicon has the largest impact on performance. Finally, as part of the contribution of this investigation, we present a novel MSA data set annotated for SSA layered on top of the PATB data annotations that will be made available to the community at large, in addition to a large scale polarity lexicon.

References

A. Abbasi, H. Chen, and A. Salem. 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Trans. Inf. Syst.*, 26:1–34.

M. Abdul-Mageed. 2008. Online News Sites and Journalism 2.0: Reader Comments on Al Jazeera Arabic. *tripleC-Cognition, Communication, Cooperation*, 6(2):59.

A. Banfield. 1982. *Unspeakable Sentences: Narration*

and Representation in the Language of Fiction. Routledge Kegan Paul, Boston.

R. Bruce and J. Wiebe. 1999. Recognizing subjectivity. a case study of manual tagging. *Natural Language Engineering*, 5(2).

T. Joachims. 2008. Svmlight: Support vector machine. http://svmlight.joachims.org/, Cornell University, 2008.

S. Kim and E. Hovy. 2004. Determining the sentiment of opinions. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 1367–1373.

M. Maamouri, A. Bies, T. Buckwalter, and W. Mekki. 2004. The penn arabic treebank: Building a large-scale annotated arabic corpus. In *NEMLAR Conference on Arabic Language Resources and Tools*, pages 102–109.

R. Tsarfaty, D. Seddah, Y. Goldberg, S. Kuebler, Y. Versley, M. Candito, J. Foster, I. Rehbein, and L. Tounsi. 2010. Statistical parsing of morphologically rich languages (spmrl) what, how and whither. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, Los Angeles, CA.

J. Wiebe, R. Bruce, and T. O'Hara. 1999. Development and use of a gold standard data set for subjectivity classifications. In *Proc. 37th Annual Meeting of the Assoc. for Computational Linguistics (ACL-99)*, pages 246–253, University of Maryland: ACL.

J. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin. 2004. Learning subjective language. *Computational linguistics*, 30(3):277–308.

J. Wiebe. 1994. Tracking point of view in narrative. *Computional Linguistics*, 20(2):233–287.

T. Wilson, J. Wiebe, and P. Hoffmann. 2009. Recognizing Contextual Polarity: an exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35(3):399–433.

J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack. 2003. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *Proceedings of the 3rd IEEE International Conference on Data Mining*, pages 427–434.

H. Yu and V. Hatzivassiloglou. 2003. The penn arabic treebank: Building a large-scale annotated arabic corpus. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 129–136.