Subjectivity and Sentiment Annotation of Modern Standard Arabic Newswire

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

Abstract

Subjectivity and sentiment analysis (SSA) is an area that has been witnessing a flurry of novel research. However, only few attempts have been made to build SSA systems for *morphologically-rich languages (MRL)*. In the current study, we report efforts to partially bridge this gap. We present a newly labeled corpus of Modern Standard Arabic (MSA) from the news domain manually annotated for subjectivity and domain at the sentence level. We summarize our linguistically-motivated annotation guidelines and provide examples from our corpus exemplifying the different phenomena. Throughout the paper, we discuss expression of subjectivity in natural language, combining various previously scattered insights belonging to many branches of linguistics.

1 Introduction

As the volume of web data continues to phenomenally increase, researchers are becoming more interested in mining that data and making the information therein accessible to end-users in various innovative ways. As a result, searches and processing of data beyond the limiting level of surface words are becoming increasingly important (Diab et al., 2009). The sentiment expressed in Web data specifically continues to be of high interest and value to internet users, businesses, and governmental bodies. Thus, the area of *Subjectivity and sentiment analysis (SSA)* has been witnessing a flurry of novel research. *Subjectivity* in natural language refers to aspects of language used to express opinions, feelings, evaluations, and speculations (Banfield, 1982; Wiebe, 1994) and it, thus, incorporates *sentiment*. The process of *subjectivity classification* refers to the task of classifying texts into either *Objective* (e.g., *More than 1000 tourists have visited Tahrir Square, in downtown Cairo, last week.*) or *Subjective*. Subjective text is further classified with *sentiment* or *polarity*. For sentiment classification, the task refers to identifying whether a subjective text is *positive* (e.g., *The Egyptian revolution was really impressive!*), *negative* (e.g., *The bloodbaths that took place in Tripoli were horrifying!*), *neutral* (e.g., *The company may release the software next month.*), and, sometimes, *mixed* (e.g., *I really like this labtop, but it is prohibitively expensive.*). SSA sometimes incorporates identifying the *holder(s)*, *target(s)*, and *strength (e.g., low, medium, high)* of the expressed sentiment.

In spite of the great interest in SSA, only few studies have been conducted on *morphologically-rich languages (MRL)* (i.e., languages in which significant information concerning syntactic units and relations are expressed at the word-level (Tsarfaty et al., 2010)). Arabic, Hebrew, Turkish, Czech, and Basque are examples of MRLs. SSA work on MRLs has been hampered by lack of annotated data. In the current paper we report efforts to manually annotate a corpus of Modern Standard Arabic (MSA), a morphologically-rich variety of Arabic, e.g., (Diab et al., 2007; Habash et al., 2009). The corpus is a collection of documents from the newswire genre covering several domains such as politics and sports. We label the data at the sentence level. Our annotation guidelines explicitly incorporate linguistically-motivated information.

Proceedings of the Fifth Law Workshop (LAW V), pages 110–118,
Portland, Oregon, 23-24 June 2011.

The rest of the paper is organized as follows: In Section 2, we motivate work on the news genre. In Section 3, we summarize our linguistically-motivated annotation guidelines. In Section 4, we introduce the domain annotation task. In Section 5 we provide examples from our dataset. We present related work in Section 6. We conclude in Section 7.

2 Subjectivity and Sentiment in the News

Most work on SSA has been conducted on data belonging to highly subjective, user-generated genres such as blogs and product or movie reviews where authors express their opinions quite freely (Balahur and Steinberger, 2009). In spite of the important role news play in our lives (e.g., as an influencer of the social construction of reality (Fowler, 1991), (Chouliaraki and Fairclough, 1999), (Wodak and Meyer, 2009)), the news genre has received much less attention within the SSA community. This role of news and the connection between news-making and social contexts and practices motivates the task of building SSA system. In addition, the many novel ways online news-making is becoming an interactive process (Abdul-Mageed, 2008) further motivates investigating the newswire genre. News-makers reproduce some of the views of their readers (e.g., by quoting them) and they devote full stories about the interactions of web users on social media outlets[1]. Although subjectivity in news articles has traditionally tended to be implicit, the fact that news stories have their own biases (e.g., hiding agents behind negative or positive events via use of passive voice, variation in lexical choice) has been pointed out by e.g., (Van Dijk, 1988). The growing trend to foster interactivity and more heavily report communication of internet users within the body of news articles is likely to make expression of subjectivity in news articles more explicit.

3 Subjectivity and Sentiment Annotation (SSA)

Two graduate level educated native speakers of Arabic annotated 2855 sentences from Part 1 V 3.0 of the Penn Arabic TreeBank (PATB) (Maamouri et al., 2004). 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. The task was to annotate MSA news articles at the sentence level. Each article has been processed such that coders are provided sentences to label. We prepared annotation guidelines for this SSA task focusing specifically on the newswire genre. We summarize the guidelines next, illustrating related and relevant literature.

	OBJ	**S-POS**	**S-NEG**	**S-NEUT**	**Total**
OBJ	1192	21	57	11	1281
S-POS	47	439	2	3	491
S-NEG	69	0	614	6	689
S-NEUT	115	2	9	268	394
Total	1423	462	682	288	2855

Table 1: Agreement for SSA sentences

3.1 SSA Categories

For each sentence, each annotator 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). We followed (Wiebe et al., 1999) in operationalizing the subjective vs. the objective categories. In other words, if the primary goal of a sentence is perceived to be the objective reporting of information, it was labeled OBJ. Otherwise, a sentence would be a candidate for one of the three subjective classes.[2] Table 1 shows the contingency table for the two annotators judgments. Overall agreement is 88.06%, with a Kappa (k) value of 0.38.

To illustrate, a sentence such as "The Prime Minister announced that he will visit the city, saying that he will be glad to see the injured", has two authors (the story writer and the Prime Minister indirectly quoted). Accordingly to our guidelines, this sentence should be annotated S-POS tag since the part related to the person quoted (the Prime Minis-

[1]This trend has increased especially in Arab news organizations like Al-Jazeera and Al-Arabiya with the hightened attention to social media as a result of ongoing revolutions and protests in the Arab world

[2]It is worth noting that even though some SSA researchers include subjective mixed categories, we only saw such categories attested in less than $< 0.005\%$ which is expected since our granularity level is the sentence. If we are to consider larger units of annotation, we believe mixed categories will become more frequent. Thus we decided to tag the very few subjective mixed sentences as S-NEUT.

ter) expresses a positive subjective sentiment, "glad" which is a *private state* (i.e., a state that is not subject to direct verification) (Quirk et al., 1974).

3.2 Good & Bad News

News can be good or bad. For example, whereas "Five persons were killed in a car accident" is bad news, "It is sunny and warm today in Chicago" is good news. Our coders were instructed not to consider *good* news *positive* nor bad news *negative* if they think the sentences expressing them are objectively reporting information. Thus, bad news and good news can be OBJ as is the case in both examples.

3.3 Perspective

Some sentences are written from a certain *perspective* (Lin et al., 2006) or point of view. Consider the two sentences (1) "Israeli soldiers, our heroes, are keen on protecting settlers" and (2) "Palestinian freedom fighters are willing to attack these Israeli targets". Sentence (1) is written from an Israeli perspective, while sentence (2) is written from a Palestinian perspective. The perspective from which a sentence is written interplays with how sentiment is assigned. Sentence (1) is considered positive from an Israeli perspective, yet the act of protecting settlers is considered negative from a Palestinian perspective. Similarly, attacking Israeli targets may be positive from a Palestinian vantage point, but will be negative from an Israeli perspective. Coders were instructed to assign a tag based on their understanding of the type of sentiment, if any, the author of a sentence is trying to communicate. Thus, we have tagged the sentences from the perspective of their authors. As it is easy for a human to identify the perspective of an author (Lin et al., 2006), this measure facilitated the annotation task. Thus, knowing that the sentence (1) is written from an Israeli perspective the annotator assigns it a S-POS tag.

3.4 Epistemic Modality

Epistemic modality serves to reveal how confident writers are about the truth of the ideational material they convey (Palmer, 1986). Epistemic modality is classified into *hedges* and *boosters*. *Hedges* are devices like *perhaps* and *I guess* that speakers employ to reduce the degree of liability or responsibility they might face in expressing the ideational material. *Boosters*[3] are elements like *definitely, I assure that*, and *of course* that writers or speakers use to emphasize what they really believe. Both hedges and boosters can (1) turn a given unit of analsysis from objective into subjective and (2) modify polarity (i.e., either strengthen or weaken it). Consider, for example, the sentences (1) "Gaddafi has murdered hundreds of people", (2) "Gaddafi may have murdered hundreds of people", and (3) "Unfortunately, Gaddafi has definitely murdered hundreds of people". While (1) is OBJ, since it lacks any subjectivity cues), (2) is S-NEUT because the proposition is not presented as a fact but rather is softened and hence offered as subject to counter-argument, (3) is a strong S-NEG (i.e., it is S-NEG as a result of the use of "unfortnately", and *strong* due to the use of the booster *definitely*). Our annotators were explicitly alerted to the ways epistemic modality markers interact with subjectivity.

3.5 Illocutionary Speech Acts

Occurrences of language expressing (e.g. *apologies, congratulations, praise*, etc. are referred to as *illocutionary speech acts* (ISA) (Searle, 1975). We believe that ISAs are relevant to the expression of sentiment in natural language. For example, the two categories *expressives* (e.g., congratulating, thanking, apologizing and *commisives* (e.g., promising) of (Searle, 1975)'s taxonomy of ISAs are specially relevant to SSA. In addition, (Bach and Harnish, 1979) define an ISA as a medium of communicating attitude and discuss ISAs like *banning, bidding, indicting, penalizing, assessing* and *convicting*. For example, the sentence "The army should never do that again" is a *banning* act and hence is S-NEG. Although our coders were not required to assign ISA tags to the sentences, we have brought the the concept of ISAs to their attention as we believe a good understanding of the concept facilitates annotating data for SSA.

3.6 Annotator's Background Knowledge

The type of sentiment expressed may vary based on the type of background knowledge of an annota-

[3] (Polanyi and Zaenen, 2006) call these *intensifiers*.

Domain	# of Cases
Politics	1186
Sports	530
Military & political violence	435
Disaster	228
Economy	208
Culture	78
Light news	72
Crime	62
This day in history	56
Total	**2855**

Table 2: Domains

tor/reader (Balahur and Steinberger, 2009). For example, the sentence "Secularists will be defeated", may be positive to a reader who opposes secularism. However, if the primary intention of the author is judged to be communicating negative sentiment, annotators are supposed to assign a S-NEG tag. In general, annotators have been advised to avoid interpreting the subjectivity of text based on their own economic, social, religious, cultural, etc. background knowledge.

4 Domain Annotation

The same two annotators also manually assigned each sentence a domain label. The domain labels are from the news genre and are adopted from (Abdul-Mageed, 2008). The set of domain labels is as follows: {*Light news, Military and political violence, Sport, Politics, Crime, Economy, Disaster, Arts and culture, This day in history*}. Table 2 illustrates the number of sentences deemed for each domain. Domain annotation is an easier task than subjectivity annotation. Inter-annotator agreement for domain label assignment is at 97%. The two coders discussed differences and a total agreement was eventually reached. Coders disagreed most on cases belonging to the *Military and political violence* and *Politics* domains. For example, the following is a case where the two raters disagreed (and which was eventutally assigned a *Military and political violence* domain):

طلب رئيس الوزراء السابق في جزر فيدجي ماهندرا شودري الذي أطيح به في ١٩ أيار مايو إثر حركة انقلابية، اليوم السبت باعادة حكومته إلى السلطة.

Transliteration: Tlb r}ys AlwzrA' AlsAbq fy jzr fydjy mAhndrA $wdry Al*y OTyH bh fy 19 OyAr mAyw Ivr Hrkp AnqlAbyp, Alywm Alsbt bIEAdp Hkwmth IlY AlslTp.
English: Former Prime Minister of Fiji Mahendra Chaudhry, who was ousted in May 19 after a revolutionary movement, asked on Saturday to return to office.

5 Examples of SSA categories from MSA news

We illustrate examples of each category in our annotation scheme. We also show and discuss examples for each category where the annotators differed in their annotations. Importantly, the two annotators discussed and adjudicated together the differences.

5.1 Objective Sentences

Sentences where no opinion, sentiment, speculation, etc. is expressed are tagged as OBJ. Typically such sentences relay factual information, potentially expressed by an official source, like examples 1-3 below:
(1)

(١) ويبلغ عدد المشردين في كنتية لوس انجلس نحو ٨٤ الف شخص.

Transliteration:[4] wyblg Edd Alm$rdyn fy kwntyp lws Onjlys nHw 84 Olf $xS.

English: The number of homeless in Los Angeles County is about 48 thousand.

(٢) طهران ٧-١٥ (أف ب) - وقع ١٦ انفجارا مساء اليوم السبت في وزارة الاستخبارات حيث استدعيت العديد من سيارات الاسعاف كما أكد شاهد عيان لوكالة فرانس برس.

Transliteration: ThrAn 15-7 (A f b) - wqE 16 AnfjArA msA' Alywm Alsbt fy wzArp AlAstxbArAt Hyv. AstdEyt AlEdyd mn syArAt AlIsEAf kmA Okd $Ahd EyAn lwkAlp frAns brs.

[4]We use here Buckwalter transliteration www.qamus.org.

English: Tehran 15-7 (AFP) - An eye witness affirmed to AFP that 16 explosions occurred late Saturday at the Ministry of Intelligence where many ambulances were summoned.

(٣) أعلن السائق الأيرلندي أيدي أيرفاين (جاغوار) انسحابه من سباق جائزة النمسا الكبرى.

Transliteration: AEln AlsA}q AlIyrlndy Iydy IyrfAyn (jAgwAr) {nsHAbh mn sbAq jA}zp AlnmsA AlkbrY.
English: The Irish driver Eddie Irvine (Jaguar) announced his withdrawal from the Austrian Grand Prix.

Examples 1-3 show that objective sentences can have some implicitly negative words/phrases like انسحاب "{nsHAb" ("withdrawl"). In addition, although these 3 examples convey *bad* news, they are annotated with an OBJ tag since the sentences are judged as facts, although one annotator did initially tag example 1 as S-NEG before it was resolved later. In a similar vein, the OBJ tag was also assigned to *good* news as in example 4 below:

(٤) وتوءكد أولغا صاحبة المجمع أن كل شيء ينتج محليا باستثناء الطحين والسكر والمشروبات التي يتم شراوءها من السوق.

Transliteration wtWkd AwlgA SAHbp AlmjmE An kl $y' yntj mHlyA b{stvnA' AlTHyn wAlskr wAlm$rwbAt Alty ytm $rAWhA mn Alswq.

English: Olga, the owner of the restaurant, asserts that everything is produced locally except flour, sugar and beverages, which are purchased from the market.
The OBJ tag was also assigned to sentences which are neither *good* nor *bad* news, as example 5 below:

(٥) وسبق لكامبوس الذي كان يشرف على الريان القطري في الموسم الماضي أن درب الشباب في مطلع التسعينيات.

Transliteration: wsbq lkAmbws Al*y kAn y$rf ElY AlryAn AlqTry fy Almwsm AlmADy On drb Al$bAb fy mTlE AltsEynyAt.

English: Previously, Campos, who acted as the coach of Al Rayyan in Qatar last season, coached Al Shabab in the early nineties.

5.2 Subjective Positive Sentences

Sentences that were assigned a S-POS tag included ones with positive *private states* (Quirk et al., 1974) (i.e., states that are not subject to verification). Examples 6 and 7 below are cases in point where the phrase انتعشت الآمال "AntE$t Al—mAl" ("hopes revived") and the word اطمئنان "TmnAn" ("relief") stand for unverifiable private states:

(٦) وانتعشت الآمال بالافراج عن الرهائن في الساعات الـ ٢٤ الأخيرة مع تدخل ليبيا.

Transliteration: wAntE$t Al—mAl bAlIfrAj En AlrhA}n fy AlsAEAt Al 24 AlAxyrp mE tdxl lybyA.
English: Hopes for the release of hostages revived in the last 24 hours with the intervention of Libya.

(٧) و أبدى صلات حسن اطمئنانه إلى عودة النظام والاستقرار إلى بلاده.

Transliteration: wAbdY SlAt Hsn TmnAnh IlY Ewdp AlnZAm wAlstqrAr IlY blAdh.
English: Silaat Hasan expressed relief for the return of order and stability to his country.

The subtle nature of subjectivity as expressed in the news genre is reflected in some of the positive examples, especially in directly or indirectly quoted content when quoted people express their emotion or support their cause (via e.g., using modifiers). For instance, the use of the phrases من أجل نهضة الصومال \"\" "mn Ajl nhDp Al-SwmAl" ("for the advancement of Somalia") and إلى الأبد "IlY AlAbd" ("for ever") in examples 8 and 9, respectively, below turn what would have otherwise been OBJ sentences into S-POS sentences. Again, one annotator initially tagged example 8 as OBJ).:

(٨) دعا الرئيس الصومالي مساء أمس السبت الدول المانحة وخصوصا أعضاء الجامعة العربية والاتحاد

الأوروبي إلى تقديم مساعدات إلى بلاده "\من أجل نهضة الصومال "\.

Transliteration: dEA Alr}ys AlSwmAly msA' Ams Alsbt Aldwl AlmAnHp wxSwSA AEDA' Al-jAmEp AlErbyp wAl{tHAd AlAwrwby IlY tqdym msAEdAt IlY blAdh "mn Ajl nhDp AlSwmAl".

English: The Somali President, on Saturday evening, called on the donor countries, especially members of the Arab League and the European Union, to provide assistance to his country "for the advancement of Somalia".

(٩) وأكد [الرئيس] أن صفحة الحرب الأهلية قد أتهت إلى الأبد، ويعود ذلك بشكل أساسي إلى انتهاء التدخلات الخارجية.

Transliteration: wAkd [Alrys] An SfHp AlHrb AlAhlyp qd Antht IlY AlAbd, wyEwd *'lk b$kl AsAsy IlY AnthA' AltdxlAt AlxArjyp.

English: He [The president] affirmed that was over for ever mainly because of the end of foreign/external interference.

Quoted content sometimes was in the form of *speech acts* (Searle, 1975). For example, (10) is an *expressive speech act* where the quoted person is thanking another party:

(١٠) [وأضاف:] "\شكرا من أعماق قلبي لهذا الشرف الذي يمتد مدى الحياة. "\

Transliteration: [wADAf:] "$krA mn AEmAq qlby lh'*A Al$rf Al*y ymtd mdY AlHyAp".

English: [He added:] Thank you from all my heart for this life-long honor.

Positive content was also sometimes explicitly expressed in the text belonging to the story author, especially in stories belonging to the Sports domain as is shown in (11).

(١١) ويمكن اعتبار ماتشالا (٥٠ عاما) من أنجح المدربين في القارة الآسيوية وتحديدا في منطقة الخليج، ويكفي أنه قاد المنتخب الكويتي إلى أحراز لقب كأس الخليج مرتين متّاليتين عام ٩٦ وعام ٩٨.

Transliteration: wymkn AEtbAr mAt$AlA (50 EAmA) mn AnjH Almdrbyn fy AlqArp AlAsywyp wtHdydA fy mnTqp Alxlyj, wykfy Anh qAd Almntxb Alkwyty IlY IHrAz lqb kAs Alxlyj mrtyn mttAlytyn EAmy 96 w 98

English: Máčala, 50 years old, is one of the most successful coaches in Asia, more specifically in the Gulf area, and it is enough that he lead the Kuwaiti team to winning the Gulf Cup twice in a row in 96 and 98.

5.3 Subjective Negative Sentences

Again, the more explicit negative content was found to be frequent in sentences with quoted content (as is illustrated in examples 12-14). (12) shows how the S-NEG S-POS sentiment can be very strong as is illustrated by the use of the noun phrase إصرار شيطاني "ISrAr $yTAny" ("diabolical insistence"):

(١٢) ورد أحد محامي أندريوتي جيواكينو على قرار النيابة في باليرمو واصفا إياه بأنه "\إصرار شيطاني "\ من قبل الاتهام.

Transliteration: wrd AHd mHAmy Andrywty jywAkynw sbAky ElY qrAr AlnyAbp fy bAlyrmw wASfA IyAh bAnh "ISrAr $yTAny" mn qbl AlAthAm.

English: One of lawyers of Andreotti Jjoaquino responded to the prosecutor's decision in Palermo, describing it as a "diabolical insistence" on the acusser's part.

(13) shows how political parties express their political stance toward events via use of private state expressions (e.g., بقلق كبير "bqlq kbyr" ["with great concern"]).

(١٣) وأوضح بيان لوزارة الخارجية التركية أن: "\تركيا تتابع بقلق كبير هجمات الارهابيّين التي حدثت في الأيام الأخيرة في أوزبكستان وقرغيزستان "\.

Transliteration: wAwDH byAn l- wzArp AlxArjyp Altrkyp An "trkyA ttAbE bqlq kbyr hjmAt AlArhAbyyn Alty Hdvt fy AlAyAm AlAxyrp fy AwzbkstAn wqrgyzstAn".

English: A statement from the Turkish Foreign Ministry indicated that "Turkey follows with great concern the terrorist attacks that have occurred in recent days in Uzbekistan and Kyrgyzstan".

Speech acts have also been used to express negative sentiment. For example, (14) is a direct quotation where a political figure denounces the acts of hearers. The speech act is intensified through the use of the adverb حتى "HtY" ("even"):

(١٤) وقال شارون من منصة الكنيست متوجها إلى نواب حزب العمل: "\لقد تخليتم حتى عن القسم الأكبر من المدينة القديمة "\.

Buckwalter: wqAl $Arwn mn mnSp Alknyst mtwjhA AlY nwAb Hzb AlEml "lqd txlytm HtY En Alqsm AlAkbr mn Almdynp Alqdymp."

English: Sharon, addressing Labour MPs from the Knesset, said: "You have even abandoned the biggest part of the old city".

Majority of the sentences pertaining to the *military and political violence* domain were OBJ, however, some of the sentences belonging to this specific domain were annotated S-NEG. News reporting is supposed to be objective, story authors sometimes used very negative modifiers, sometimes metaphorically as is indicated in (15). Example 15, however, was labeled OBJ by one of the annotators, and later agrrement was reached that it is more of an S-NEG case.

(١٥) وكان شهر تموز (يوليو) دمويا بشكل خاص مع سقوط نحو ٣٠٠ قتيل.

Transliteration: wkAn $hr tmwz ywlyw dmwyA b$kl xAS mE sqwT nHw 300 qtyl.
English: The month of July was especially bloody, with the killing of 300 people.

Again, authors of articles sometimes evaluated the events they reported. Sentences 16 and 17 are examples:

(١٦) وبات موقف فريق الأهلي صعبا للغاية في البطولة الأفريقية التي يسعى للفوز بلقبها وتضع جماهيره أيديها على قلوبها خشية انهياره.

Transliteration: wbAt mwqf fryq AlAhly SEbA lAlgAyp fy AlbTwlp AlIfryqyp Alty ysEY lAlfwz blqbhA wtDE jmAhyrh AydyhA ElY qlwbhA x$yp nhyArh.
English: The position of Al-Ahly in the African Championship, which the team seeks to win, became extremely difficult; and the team's fans hold their breath in fear of its defeat.

(١٧) وجاء اعتداء هشام حنفي على زميله شادي محمد على مرأى ومسمع الجميع أثناء مباراة الأهلي والاسماعيلي في نصف نهائي الكأس الأسبوع الماضي ليوءكد تفكك الفريق.

Transliteration: wjA' AEtdA' h$Am Hnfy ElY zmylh $Ady mHmd ElY mrAY wmsmE AljmyE AvnA' mbArAp AlAhly wAlAsmAEyly fy nSf nhA}y AlkAs AlAsbwE AlmADy lyWkd tfkk Alfryq.
English: Hesham Hanafi's attack on his colleague Shadi Muhammad, in front of everyone during the game between Al-Ahli and Al-Ismaili in the semi-finals last week, confirms the disintegration of the team.

5.4 Subjective Neutral Sentences

Some of the S-NEUT cases were speculations about the future, as is illustrated by sentences 18 and 19:

(١٨) ويتوقع أن يعود إلى الولايات المتحدة في ٢٥ تموز (يوليو).

Transliteration: wytwqE An yEwd IlY AlwlAyAt AlmtHdp fy 25 tmwz (ywlyw).
English: And he is expected to return to the United States on July 25.

(١٩) وكل الموءشرات تفيد أن هذا الوضع لن يتغير بعد الانتخابات.

Transliteration: wkl AlmW$rAt tfyd In h'*A AlwDE ln ytgyr bEd AlAntxAbAt.
English: All indications are that this situation will not change after the elections.

Hedges were also used to show cautious commitment to propositions, and hence turn OBJ sentences to S-NEUT ones. Sentences (20) and (21) are examples, with the occurrence of the hedge trigger word يبدو "ybdw" ("it seems") in (20) and على الأرجح "ElY AlArjH" ("it is most likely") in (21):

(٢٠) و يبدو أن التكتم الذي أحاط بزيارة بيريز إلى أندونيسيا كان يهدف إلى تفادي إثارة ردود فعل معادية في البلاد.

Transliteration: w ybdw An Altktm Al*y AHAT bzyArp byryz AlY AndwnysyA kAn yhdf AlY tfAdy AvArp rdwd fEl mEAdyp fy AlblAd.
English: It seems that the secrecy surrounding Peres's visit to Indonesia was aimed at avoiding negative reactions in the country.

(٢١) وعلى الأرجح أن قبطان الغواصة أعطى الأمر بإطفاء كل الآلات على متنها.

Transliteration: wElY AlArjH An qbTAn Al-gwASp AETY AlAmr bATfA' kl AlAlAt ElY mtnhA.
English: Most likely the submarine's captain ordered turning off all the machines on board.

Some S-NEUT cases are examples of *arguing* that something is true or should be done (Somasundaran et al., 2007). (22) is an illustrative example:

(٢٢) قلتها، وأكررها، فالمشكلة ليست في النفط الخام وإنما في المشتقات النفطية.

Transliteration: qlthA, wAkrrhA, fAlm$klp lyst fy AlnfT AlxAm wInmA fy Alm$tqAt AlnfTyp.
English: I said, and I repeat it, the problem is not in crude oil but rather in oil derivatives.

Example 22 was, however, initially tagged as OBJ. Later, the two annotators agreed to assign it an S-NEUT tag.

6 Related Work

There are a number of datasets annotated for SSA. Most relevant to us is work on the news genre. (Wiebe et al., 2005) describe a fine-grained news corpus manually labeled for SSA[5] at the word and phrase levels. Their annotation scheme involves identifying the *source* and *target* of sentiment as well as other related properties (e.g., the *intensity* of expressed sentiment). Our work is less fine grained on the one hand, but we label our data for domain as well as subjectivity.

(Balahur et al., 2009) report work on labeling quotations from the news involving one person mentioning another entity and maintain that quotations typically contain more sentiment expressions than other parts of news articles. Our work is different from that of (Balahur et al., 2009) in that we label all sentences regardless whether they include quotations or not. (Balahur et al., 2009) found that entities mentioned in quotations are not necessarily the target of the sentiment, and hence we believe that SSA systems built for news are better if they focus on all the sentences of articles rather than quotations alone (since the target of sentiment may be outside the scope of a quotation, but within that of the sentence to which a quotation belongs)..

The only work on Arabic SSA we are aware of is that of Abbasi et al. (2008) who briefly describe labeling a collection of documents from Arabic Web forums. (Abbasi et al., 2008)'s dataset, however, is not publicly available and detailed information as to how the data was annotated is lacking. Our work is different from (Abbasi et al., 2008)'s in that we label instances at the sentence level. We believe that documents contain mixtures of OBJ and SUBJ cases and hence sentence-level annotation is more fine-grained. In addition, (Abbasi et al., 2008) focus on a specific domain of 'dark Web forums'.

7 Conclusion

In this paper, we present a novel annotation layer of SSA to an already labeled MSA data set, the PATB Part 1 ver. 3.0. To the best of our knowledge, this layer of annotation is the first of its kind on MSA data of the newswire genre. We will make that collection available to the community at large. We motivate SSA for news and summarize our linguistics-motivated guidelines for data annotation and provide examples from our data set.

[5]They use the term *private states* (Quirk et al., 1974) to refer to expressions of subjectivity.

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.

K. Bach and R.M. Harnish. 1979. Linguistic communication and speech acts.

A. Balahur and R. Steinberger. 2009. Rethinking Sentiment Analysis in the News: from Theory to Practice and back. *Proceeding of WOMSA*.

A. Balahur, R. Steinberger, E. van der Goot, B. Pouliquen, and M. Kabadjov. 2009. Opinion mining on newspaper quotations. In *2009 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pages 523–526. IEEE.

A. Banfield. 1982. *Unspeakable Sentences: Narration and Representation in the Language of Fiction*. Routledge Kegan Paul, Boston.

L. Chouliaraki and N. Fairclough. 1999. *Discourse in late modernity: Rethinking critical discourse analysis*. Edinburgh Univ Pr.

M. Diab, K. Hacioglu, and D. Jurafsky. 2007. Automatic processing of Modern Standard Arabic text. *Arabic Computational Morphology*, pages 159–179.

M.T. Diab, L. Levin, T. Mitamura, O. Rambow, V. Prabhakaran, and W. Guo. 2009. Committed belief annotation and tagging. In *Proceedings of the Third Linguistic Annotation Workshop*, pages 68–73. Association for Computational Linguistics.

R. Fowler. 1991. *Language in the News: Discourse and Ideology in the Press*. Routledge.

N. Habash, O. Rambow, and R. Roth. 2009. Mada+tokan: A toolkit for arabic tokenization, diacritization, morphological disambiguation, pos tagging, stemming and lemmatization. In *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR)*, Cairo, Egypt.

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

W.H. Lin, T. Wilson, J. Wiebe, and A. Hauptmann. 2006. Which side are you on?: identifying perspectives at the document and sentence levels. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pages 109–116. Association for Computational Linguistics.

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.

F. Palmer. 1986. *Mood and Modality. 1986*. Cambridge: Cambridge University Press.

L. Polanyi and A. Zaenen. 2006. Contextual valence shifters. *Computing attitude and affect in text: Theory and applications*, pages 1–10.

R. Quirk, S. Greenbaum, R.A. Close, and R. Quirk. 1974. *A university grammar of English*, volume 1985. Longman.

J.R. Searle. 1975. A taxonomy of speech acts. In K. Gunderson, editor, *Language, mind, and knowledge*, pages 344–369. Minneapolis: University of Minnesota Press.

S. Somasundaran, J. Ruppenhofer, and J. Wiebe. 2007. Detecting arguing and sentiment in meetings. In *Proceedings of the SIGdial Workshop on Discourse and Dialogue*, volume 6. Citeseer.

H. Tanev. 2007. Unsupervised learning of social networks from a multiple-source news corpus. *MuLTISOuRcE, MuLTILINguAL INfORMATION ExTRAcTION ANd SuMMARIzATION*, page 33.

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.

T.A. Van Dijk. 1988. *News as discourse*. Lawrence Erlbaum Associates.

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, and C. Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2):165–210.

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

R. Wodak and M. Meyer. 2009. Critical discourse analysis: History, agenda, theory and methodology. *Methods of critical discourse analysis*, pages 1–33.