Multi-entity Sentiment Scoring

Karo Moilanen and Stephen Pulman
Oxford University Computing Laboratory
{ *Karo.Moilanen* | *Stephen.Pulman* }@*comlab.ox.ac.uk*

Abstract

We present a compositional framework for modelling entity-level sentiment (sub)contexts, and demonstrate how holistic multi-entity polarity scoring emerges as a by-product of compositional sentiment parsing. A data set of five annotators' multi-entity judgements is presented, and a human ceiling is established for the challenging new task. The accuracy of an initial implementation, which includes both supervised learning and heuristic distance-based scoring methods, is 5.6∼6.8 points below the human ceiling amongst sentences and 8.1∼8.7 points amongst phrases.

Keywords

Entity-level sentiment analysis, sentiment scoring, sentiment parsing, sentiment annotation, compositional semantics

1 Introduction

The ability to detect author sentiment towards various entities in text is a fundamental goal in sentiment analysis, and holds great promise for many applications. Entities, which can comprise anything from mentions of people or organisations to concrete or even abstract objects, condition what a text is ultimately about. Besides the intrinsic value of entity scoring, the success of document- and sentence-level analysis is also decided by how accurately entities in them can be modelled. Deep entity analysis unfortunately presents the most difficult challenges, be they linguistic or computational. One of the most recent developments in the area - *compositional semantics* - has shown potential for sentence- and expression-level analysis in both logic-oriented [11],[9] and machine learning-oriented [3] paradigms. Our goal in this paper is to further that avenue by extending it to entity-level sentiment analysis.

Entity-level approaches have so far involved relatively shallow methods which usually presuppose some pre-given topic or entity of relevance to be classified or scored (§5.3). Other proposals have attempted specific semantic sentiment roles such as evident sentiment HOLDERs, SOURCEs, TARGETs, or EXPERIENCERs (§5.2). What characterises these approaches is that only a few specific entities in text are analysed while all others are left unanalysed. While shallow approaches can capture some amount of explicitly expressed sentiment, they ignore all layers of implicit sentiment pertaining to a multitude of other entities. We believe that access to these rich layers is required for deeper logical sentiment reasoning in the future.

We take a different view on the problem and investigate the possibility of a holistic *multi*-entity analysis in that we make no categorical distinctions between individual entity mentions, topics, or sentiment roles of any kind. We instead refer to all base nouns simply as **entity markers** which may (or may not) serve the above metafunctions, and aim at classifying *all* such markers in sentences using a single, unified approach. For the sentence in Ex. 1, we envisage a classifier that classifies all of the bracketed entities as positive$^{(+)}$, neutral$^{(N)}$, or negative$^{(-)}$ (NB. / = 'or'):

(1) *"Here's the* `[thing]`$^{(N)/(+)}$*: Other* `[studies]`$^{(N)/(+)}$ *have found that* `[clergy]`$^{(+)}$*, and not* `[psychologists]`$^{(-)/(+)}$ *or other mental* `[health]`$^{(+)}$ `[experts]`$^{(+)/(-)}$*, are the most common* `[source]`$^{(+)/(N)}$ *of* `[help]`$^{(+)}$ *sought in* `[times]`$^{(N)/(-)}$ *of psychological* `[distress]`$^{(-)}$*."*

Note that, in this kind of deep analysis, not only can the polarity of an entity differ from the global, sentential reading but it may also depend heavily on one's subjective point of view: for example, the entity `[experts]` is logically either positive or negative, arguably. Simple keyword spotting, window-based techniques, and even statistical features have limited power in multi-entity analysis because of the inherently overlapping and interdependent nature of entities. We argue in this paper that the analytical strategy towards this problem needs to be grammatical in nature.

Going beyond existing shallow single-entity approaches to deep multi-entity scoring requires the 'conventional' definitional scope of sentiment to be extended to include not only 1) explicit subjective expressions of sentiment, opinions, and emotions, but also 2) implicit subjective expressions and connotations describing some positive (desirable, favourable), negative (undesirable, unfavourable), or neutral (objective) state of affairs in the world. Our classification task is accordingly much wider than most past work in the area. We now illustrate how existing compositional approaches can be extended for multi-entity scoring purposes.

2 Sentiment Parsing

We adopted the compositional sentiment model described in [11] as a basis for our scoring framework. In *idem.*, polarity classification is broken down into binary combinatory steps whereby two syntactic input (`IN`) constituents are combined at a time, and a three-valued polarity logic controlled by a sentiment grammar calculates a polarity for the resultant composite constituent. The process starts with word-level lexical

International Conference RANLP 2009 - Borovets, Bulgaria, pages 258–263

Table 1: *Sample Constituent Rankings*

Mod:AdjP	≫	Head:N	*[funny blunders]*$^{(+)}$
Mod:Nom	≪	Head:N	*[error reduction]*$^{(+)}$
Mod:AdvP	≫	Head:Adj	*[badly decorated]*$^{(-)}$
Head:Adj	≫	Comp:PP	*[sick of fame]*$^{(-)}$
Head:N	≪	Comp:VP	*[market gone sour]*$^{(-)}$
Head:Pred	≫	Comp:DirObj	*[end the hostility]*$^{(+)}$
Head:Pred	≪	Adjunct:Adv	*[smiled painfully]*$^{(-)}$
...			

seeds, proceeds recursively via intermediate syntactic levels, and terminates at the top sentence level.

2.1 Compositional Processes. The model in *idem.* operates with positive (POS), negative (NEG), and neutral (NTR) polarities, and reversive (¬) and equative (=) polarity shifting values. Non-neutral sentiment propagation is modelled by allowing non-neutral (POS, NEG) constituents to override NTR ones (e.g. *"[funny*$^{(+)}$ *things*$^{(N)}$*]*$^{(+)}$*"*). The model supports polarity-reversing compositions (cf. [14]) in which reversive (¬) constituents reverse non-neutral ones (e.g. *"[no*$^{[\neg]}$*talent*$^{(+)}$*]*$^{(-)}$*"*; *"[tax*$^{(-)}$ *decreases*$^{[\neg]}$*]*$^{(+)}$*"*), and the resolution of non-neutral polarity conflicts (e.g. *"[bad*$^{(-)}$*luck*$^{(+)}$*]*$^{(-)}$*"*; *"[cancer*$^{(-)}$ *cure*$^{(+)}$*]*$^{(+)}$*"*).

2.2 Sentiment Grammar. Since the polarity of a composite constituent can differ from the two IN polarities, the IN constituents can not be equally salient. The model assigns relative weights to the two IN constituents to dictate whose sentiment dominates: the stronger of the two (superordinate (SPR)) dominates the weaker one (subordinate (SUB)) (i.e. SPR ≫ SUB). The weights are not stored in any individual IN constituents but are latent in specific syntactic constructions such as [Mod:Adj Head:N] (i.e. adjectival premodification of head nouns) or [Head:V Comp:NP] (i.e. direct object complements of verbs). Crucially then, a constituent may be superordinate in one syntactic environment but subordinate somewhere else: consider *"helpline*$^{(+)}$*"* in *"[abuse helpline]*$^{(+)}$*"* vs. *"[useless helpline]*$^{(-)}$*"*, for example. The effects of different syntactic environments on IN constituent rankings are specified in a hand-written sentiment grammar which is described in more detail in [11]. Table 1 illustrates some sample grammatical rankings.

2.3 Pre-processing. Raw text is first processed with a dependency parser[1]. A flat parse tree is then generated in which each constituent head is linked to zero or more pre- and/or post-head dependents. Each leaf node is assigned a prior sentiment polarity and reversal value. These are obtained from an extensive word-class-specific, general-purpose main sentiment lexicon of 57103 sentiment words (22402 ADJ, 6487 ADV, 19004 N, 9210 V), and from an auxiliary list of 312343 known NTR words. Our main lexicon, which was compiled manually based on WordNet 2.1 synsets and glosses, contains 21341 POS, 7036 NTR, and 28726 NEG entries; 1700 (3%) have (¬) reversal features.

2.4 Parsing. Sentiment analysis starts with the main lexical head verb of the root clause (or the head noun of a main clausal NP), and first descends recursively down to its lowermost atomic leaf constituents. Through a recursive bottom-up traversal of the dependency tree, each constituent's internal polarity is resolved before it is combined with its parent constituent. When parsing a constituent, the parser follows a fixed order in combining the constituent head (H_i) first with j post-head ($R_{i+1\,:\,i+j}$) dependents and then with k pre-head ($L_{i-k\,:\,i-1}$) dependents (schematised in Fig. 1). Each combinatory step operates on the head and only one of its dependents, and consults the sentiment grammar (§2.2) to determine which element is SPR and assigns the resultant compositional polarity to the head-dependent pair.

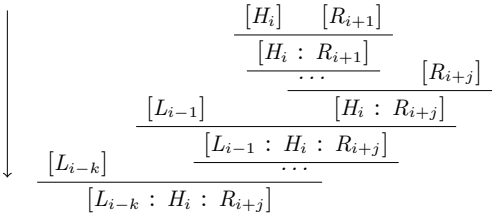


Fig. 1: *Head-dependents combination schema*

3 Entity Scoring

Since each constituent - a head with k pre- j post-head dependents - stands for a unique (sub)part of the sentence (i.e. $[L_{i-k} : H_i : R_{i+j}]$), a constituent and its internal polarity constitutes a **sentiment** (**sub**)**context** in the sentence. Each constituent consequently shapes the polarities of the entity marker(s) inside it. Leaf-node (sub)contexts holding but a single entity marker can be seen as intrinsically **lexical** for they represent atomic pieces of information without alluding to any higher context(s). In contrast, (sub)contexts in which entity markers fall under the influence of other words are extrinsically **contextual**. Importantly then, the very possibility of expressing opinions and sentiments about an entity means that a sentence can exhibit many contextual polarities for it. These can and often do differ from the atomic lexical polarity of the entity and the polarity of the sentence. In the headline *"[EU opposes* [credit] *crunch rescue package]*$^{(-)}$*"*, the entity [credit] is shaped by six (sub)contexts (Ex. 2):

(2)
1: [[credit]]$^{(+)}$
2: [[credit] *crunch*]$^{(-)}$
3: [[credit] *crunch rescue*]$^{(+)}$
4: [[credit] *crunch rescue package*]$^{(+)}$
5: [*opposes* [credit] *crunch rescue package*]$^{(-)}$
6: [*EU opposes* [credit] *crunch rescue package*]$^{(-)}$

We aim at including in our analysis not only the two extremes (1: atomic lexical, 6: global sentential) but all intermediate levels of sentiment as well. Seen as a stack of (sub)contexts, the occurrences of an entity across all (sub)contexts along the atomic-global continuum give rise to three gradient polarity distribution scores (#POS, #NTR, #NEG). Entity-level sentiment scoring thus involves measuring how many times each entity was found in POS, NTR, and NEG (sub)contexts. The scoring process is incremental in that each time the parser has calculated a compositional polarity for a constituent (i.e. a (sub)context), we locate all entity markers inside the (sub)context, and, for each found entity marker, use the polarity distribution within the

[1] Connexor Machinese Syntax (www.connexor.com)

(sub)context to increment the entity's polarity counts, accordingly.

The main challenge is how (sub)contexts' polarity distributions are actually measured. We experimented with two possible scoring methods. Our scoring framework is however not restricted to any particular scoring method(s) per se as other scorers can be plugged in.

3.1 Distance Scoring. The most basic method for measuring the polarity distribution of a (sub)context is a bidirectional polarity search around an entity marker word. For polarity $p \in \{\texttt{POS}, \texttt{NTR}, \texttt{NEG}\}$, in a (sub)context with n neighbouring words with p around an entity marker word at word ID w_m, the following distance scoring function is used within each (sub)context:

$$dist(p) = \sum_{i=1}^{n} \frac{1}{worddist(w_m, w_i^p)} \cdot \frac{\Theta}{clausedist(w_m, w_i^p)}$$

In addition to the raw distance between the entity marker and a neighbouring word (*worddist*), the distance between their respective (full) clause IDs is also considered (*clausedist*). The Θ coefficient, which was set experimentally at 1.75, boosts neighbouring words that are in the same (full) clause as the entity marker. Because only some higher-level (sub)contexts contain subregions with contrasting polarities (e.g. multiple clauses), distance scoring often suggests similar polarity distributions for all entities in a given (sub)context.

3.2 Syntactic Scoring. Distance scoring takes no notice of syntactic or lexical evidence around entity markers. Such blanket coverage risks being too broad. For more complex scoring, we used supervised learning with Support Vector Machines[2]. We apply the feature template in Table 2 to ± 3 words around each entity marker (within a (sub)context). The PRIOR_POLARITY and POLARITY_REVERSAL features refer to a word's raw prior lexical polarity and polarity reversal values while GLOBAL_POLARITY indicates the current (sub)context's internal polarity (as suggested by the parser). The DEPENDENCY_TYPE, GRAMMATICAL_RELATION, SYNTACTIC_ROLE, and WORD_CLASS features reflect the tags assigned to each word by the dependency parser. POLARITY_WSD_TYPE indicates whether a word is tagged in the lexicon as capable of bearing more than one polarity (e.g. *"lean*$^{(N)(+)(-)}$*"*, *"chicken*$^{(N)(-)}$*"*, *"bliss*$^{(+)}$*"*). UNIGRAM features are also included. In total, 19502 binary features (§4.1) were used to train a polynomial kernel.

Based on the observed variability in human annotations in the training data (§4.1), we trained five separate models (one per annotator), and run them as a committee. In each (sub)context, each entity marker word is submitted to the committee and the number of classifiers returning polarity $p \in \{\texttt{POS}, \texttt{NTR}, \texttt{NEG}\}$ as a class label is used to increment the entity's corresponding polarity counts:

$$svmvote(p) = \#\text{ of SVMs classifying entity as } p$$

SVMs' predicted class labels are required to fulfill one post-classification polarity axiom: if a (sub)context does not contain any words with POS or NEG prior polarities (i.e. it is fully NTR), non-neutral predictions are discarded and asserted as NTR instead.

[2] Johnson, M. (2008). SVM.NET 1.4. (www.matthewajohnson.org/software/svm.html). Based on Chang, C. & Lin, C. (2001). LIBSVM. (www.csie.ntu.edu.tw/~cjlin/libsvm/).

Table 2: *SVM entity feature template*

PRIOR_POLARITY	GLOBAL_POLARITY	UNIGRAM
POLARITY_REVERSAL	POLARITY_WSD_TYPE	
DEPENDENCY_TYPE	SYNTACTIC_ROLE	
GRAMMATICAL_RELATION	WORD_CLASS	

3.3 Weights. The sentiment parsing process scores entities incrementally by measuring the polarity distribution of one (sub)context at a time and updating the entities in it. The cumulative polarity distributions $D_1 \ldots D_n$ of an entity across all of its hosting (sub)contexts $z_1 \ldots z_n$ ultimately determine the entity's final sentiment scores. However, simple cumulative sums do not suffice. In particular, individual (sub)contexts' scores need to be weighted because not all of them are equally salient: atomic (sub)contexts are evidently not very important, for example.

We experimented with three empirically discovered coefficients to control the weight of each (sub)context. g estimates the information gain of a (sub)context over its predecessor by boosting longer (sub)contexts. β measures the length of a (sub)context in the sentence: longer (sub)contexts are again boosted. Abrupt polarity changes between (sub)contexts are boosted by v: for example, a NEG (sub)context followed by a POS one may indicate a shift in perspective or negation. For each entity, the cumulative score for polarity $p \in \{\texttt{POS}, \texttt{NTR}, \texttt{NEG}\}$ in a sentence with n (sub)contexts $(z_1 \ldots z_n)$ is obtained as follows:

$$scr(p) = \sum_{i=1}^{n} \frac{g\beta v D_i}{len(z_i)}$$

D_i = $dist(p)$ or $svmvote(p)$ score from (sub)context z_i
g = $length(z_i)$ - $length(z_{i-1})$
β = $length(z_i)$ / $length(sentence)$
v = 1.75 if polarity of z_i is not polarity of z_{i-1}, else 1

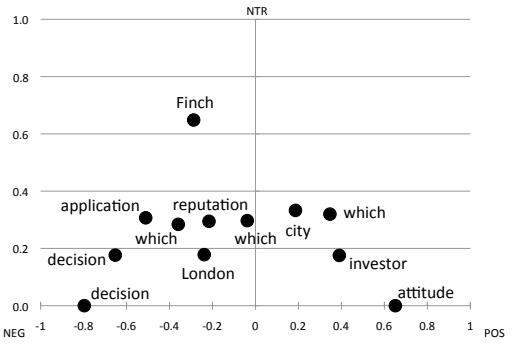
3.4 Sample Analysis. Consider Ex. 3:

(3) *"**Finch** said the **decision** to withdraw the **application** was a 'dispiriting **decision which** will harm **London's reputation** as a **city which** is well governed, and **which** hitherto has had a welcoming **attitude** to major overseas **investors**'."*

Since the sentence depicts a state of affairs that is negative/undesirable/unfavourable, all entities in it could be classified uniformly as NEG. However, the sentential negativity *does not entail* that *"[a city which is well governed]*$^{(+)}$*"* and *"[a welcoming attitude to major overseas investors]*$^{(+)}$*"* are NEG as such: instead, it merely makes an allusion to their involvement in a NEG context. The same holds for *"[London's reputation]*$^{(N)(+)}$*"*. We therefore expect the algorithm to assign different degrees of negativity (and positivity and neutrality) to the entities. Ex. 4 visualises the parser's entity scores. The polarity scores of the entity [London] (29% POS, 18% NTR, 53% NEG), which are illustrated in Table 3, reflect the statement in that (i) [London] is NTR in itself, (ii) it has a positively-evaluated reputation, and (iii) it is affected by a NEG event. The other entities, from the most NEG [decision] to the most NTR [Finch], are tenable, too. Note that the scores represent each entity's involvement in three polarity contexts and may not

as such indicate sentiment/polarity strength although small margins amongst the three values signal mixed (sub)contexts while large(r) margins can be equated with pure(r) polarities. We observed that interpreting these kinds of multi-entity scores is similar to interpreting automatically generated summaries in that, due to subjective scaling and class in- and exclusion preferences, the scores often afford *many* possible interpretations: whether the NEG score for [application] should be .82, SOMEWHAT_NEG, or some other arbitrary value, for example, is secondary to the fact that the parser ranked the entity sensibly as NEG ≫ NTR ≫ POS.

(4)



4 Evaluation

4.1 Gold Standard. Most existing gold standards used in past sentiment research such as MPQA[3] [20] or FBS[4] [4] come with incomplete entity annotations as only some entities (e.g. sentiment roles or product features) are usually included per text region. In contrast, we wish to evaluate *all* entity markers in a given text region. To achieve that, a new multi-entity data set was compiled from a cross-genre pool of 24 documents' dependency parses. Five annotators (three paid linguistics students, one of the authors, one volunteer) annotated 7904 entity markers as POS, NTR, or NEG (cf. Ex. 1). Cases displaying mixed sentiment or those infected with inescapable ambiguity were marked as ambiguous. In order to preclude misaligned annotations between annotators (cf. [20]: 34-42; [7]: 6), we made a decision to confine ourselves to base nouns only (cf. Ex. 1).

The data set contains two subsections. The first (GS_PHR) contains 4765 entities from 1500 syntactic constituent phrases of differing lengths (from six documents) while the second (GS_SNTC) encompasses 3139 entities from 500 full sentences (from 18 documents). Both subsets were further split into 4/5 training and 1/5 testing sections, yielding for training 2490 entities (GS_SNTC) vs. 3877 entities (GS_PHR) (§3.2). 649 and 888 entities are given for testing, respectively. For syntactic scoring, the SVM classifier committee consisted of five separate models, each trained on 6367 entities (with 19502 features) from one annotator's combined GS_SNTC and GS_PHR training sections (§3.2).

4.2 Human Ceiling. In order to estimate human performance in the new task, we compared each annotator against all others, and obtained five average accuracy and Kappa scores (Table 4). It is apparent that the task is highly subjective because the figures are only modest in a three-way condition (accuracy 62%; k .43∼.45) (see §4.4). However, the task is considerably less vague in a two-way non-neutral condition (86∼89% accuracy; k .70∼.78).

4.3 Error Classification. The inter-annotator agreement levels point towards increased ambiguity with NTR polarity due to differing personal degrees of sensitivity towards neutrality/objectivity. Not all classification errors are then equal for classifying a POS case as NTR is more tolerable than classifying it as NEG, for example. We found it useful to characterise three distinct error classes or disagreements between human H and algorithm A. FATAL errors ($H^{(\alpha)}A^{(\neg\alpha)}$ $\alpha\in\{+\ -\}$) are those where the non-neutral polarity is completely wrong: such errors affect the performance of the parser adversely. GREEDY errors ($H^{(N)}A^{(\alpha)}$ $\alpha\in\{+\ -\}$) are those where the algorithm wrongly made a decision to jump one way or the other, displaying oversensitivity towards non-neutral polarities. LAZY errors ($H^{(\alpha)}A^{(N)}$ $\alpha\in\{+\ -\}$) indicate that the algorithm chose to sit on the fence and displayed oversensitivity towards NTR polarity.

4.4 Test Conditions. The highest-scoring polarity (1^{st} rank) amongst each entity's three polarity counts is compared against the gold standard. All ambiguous cases were excluded, as were a few tie scores amongst short phrases. We compare the DIST and SVM scorers against a fully-COMPOSitional baseline that simply uses the internal polarity of a (sub)context to score its entities. A hybrid DIST+SVM method is also evaluated. All experiments were conducted under a (i) three-way ALL_POL (POS:NTR:NEG), and a (ii) two-way NON_NTR (POS:NEG, with FATAL errors only) classification condition. The proportions of finding a match in the algorithm's 1^{st}, 2^{nd}, and 3^{rd} polarity ranks are included. The algorithm's average figures against five annotators are given in Table 5.

4.5 Results. In absolute terms, the results are modest. But in comparison with the low human ceiling, the algorithm's best scores are only 5.6∼8.7 points behind (ALL_POL). Both scorers outperformed the fully COMPOSitional baseline - a realisation implying that entity-level sentiment is weakly compositional although, interestingly, non-compositional scoring can be approached compositionally. Shorter constituents with less contextual evidence (GS_PHR) were, as expected, more challenging than longer, holistic constituents (GS_SNTC). Most notable is the performance of the heuristic DIST method which generally equalled or outperformed the SVM committee. The hybrid combination (DIST+SVM) resulted in a small boost. The two complementary scoring methods appear to neutralise each other's errors as DIST displays oversensitivity towards POS and NEG labels (cf. more GREEDY errors) while SVM suggests NTR in many cases (cf. mostly LAZY errors). The correct label was in the parser's 1^{st} and 2^{nd} ranks in 79∼85% of the cases (ALL_POL) which confirms that the parser generally points at the right direction. Matching past observations in the area, the average gap between three-way ALL_POL and two-way NON_NTR classification accuracy is noticeable at 20∼25 points.

4.6 Future Work. Further research is needed to address cases of 'sentiment overflow' where an entity's scores are incorrectly shaped by (sub)contexts beyond its natural sentiment zone boundaries. Although en-

[3] http://www.cs.pitt.edu/mpqa/
[4] http://www.cs.uic.edu/~liub/

Table 3: *Sample analysis of (sub)contexts containing the entity "London" (with* POS:NTR:NEG *scores)*

SUBCONTEXT	TYPE	ENTITY MARKER
*[**London's**]$^{(N)}$*	Lexical	[London] 0:100:0
*[**London's** reputation]$^{(+)}$*	Contextual	[London] 5:95:0
*[which will harm **London's** reputation as a city which is well governed, and which hitherto has had a welcoming attitude to major overseas investors']$^{(-)}$*	Contextual	[London] 27:30:43
*[a 'dispiriting decision which will harm **London's** ... investors']$^{(-)}$*	Contextual	[London] 28:24:48
*[the decision to withdraw the application was ... **London's** ... investors']$^{(-)}$*	Contextual	[London] 29:20:51
*[Finch said the decision to withdraw the application was a 'dispiriting decision which will harm **London's** reputation as a city which is well governed, and which hitherto has had a welcoming attitude to major overseas investors']$^{(-)}$*	Contextual, global	[London] 29:18:53

Table 4: *Human accuracy and inter-annotator agreement scores on the gold standard*

	GS_SNTC (3139)				GS_PHR (4765)			
	k ALL_POL	k NON_NTR	Acc ALL_POL	Acc NON_NTR	k ALL_POL	k NON_NTR	Acc ALL_POL	Acc NON_NTR
Human-1	.50	**.82**	66.82	**90.99**	**.49**	**.74**	66.83	**87.90**
Human-2	.48	.77	65.03	88.67	**.49**	.71	**66.87**	86.43
Human-3	.34	.79	52.79	89.60	.33	.72	55.09	86.73
Human-4	**.51**	.80	**66.90**	89.70	.47	.66	64.46	82.88
Human-5	.40	.72	58.80	86.21	.36	.69	54.89	85.14
Avg	.45	.78	62.07	89.03	.43	.70	61.63	85.81

tity markers (and any sentiment roles therein) are linked through a variety of complex means [16][18][6], taking discourse structure, Named Entities, semantic roles, and reported speech into account would be beneficial. Entity markers can be chained through anaphora/co-reference resolution which can lead to significant boosts [6]. The values for the weighting coefficients (§3.3) and the exploratory learning features for syntactic scoring (§3.2) can be optimised, and other scorers may be employed.

5 Related Work

5.1 Compositional Analysis. A few systems that exploit the compositional properties of sentiment in differing degrees have been proposed. The system closest to our framework is [9] who describe a tool for phrase- and sentence-level classification. A sentiment composition model is described which uses a cascade of transducers relying on lexical sentiment seeds, a phrasal chunker, and hand-written pattern-matching rules. Instead of making use of compositional rules (cf. §2.2), [3] incorporated compositional semantics into structured inference-based learning with lexical, negator, and voting features. [12] describe a hybrid system for detecting sentiment expressions about a topic that combines a rule-based sentiment extractor with a learning-based topic classifier. For the former, phrasal chunking and shallow parsing patterns are used to combine elements in specific syntactic cases. However, no explicit details about compositional processes are given. [17] uses scored prior polarities from sentiment lexica and knowledge bases with dependency parsing to generate verb-centric ACTOR-ACTION-OBJECT frames (each with optional internal modifiers), and calculate contextual polarities at different structural levels using hand-written polarity combination rules. A shallow compositional affect sensing approach with lexical, phrasal, and sentential linking and ranking patterns is proposed in [13].

5.2 Entities. In classifying raw entity mentions without deep sentiment semantics, the primary focus has been on relatively shallow techniques restricted to specific topical mentions, or product names, features, and attributes. Goalwise, the approach closest to our multi-entity framework is [6] who classify entities (topics) expressed in IR search queries. Matched query entities are expanded through co-reference and meronymy analysis of concrete entities' parts and features to generate a set of topical entity mentions. These are paired with topically relevant sentiment expressions targeting them, and aggregate scores for the query entities are calculated using a sentiment propagation graph. For each sentiment expression, candidate target mentions are ranked with proximity-based, heuristic, and supervised learning-based scorers.

The product feature mining and summarisation system described in [5] classifies feature mentions based on neighbouring adjectives and sentential polarity frequencies. [4] propose a more complex approach targeting products' parts and attributes with a holistic lexicon- and distance-based method that exploits local and global clause-, sentence-, and review-level evidence and patterns in disambiguating ambiguous words, irregular/idiomatic constructions, and polarity conflicts. A relaxation labelling technique was used in [15] to classify product feature mentions by sequential analyses of words, features, and sentences with syntactic dependency, lexical, and collocational constraints. [10] extract opinions with fixed opinion frames which capture for a given entity an attribute and a sentiment expression with its HOLDER.

5.3 Sentiment Roles. The inventory of possible semantic roles *specific* to sentiment is unclear. Past proposals have targeted some of the most obvious roles encompassing opinion HOLDERs, SOURCEs, TARGETs, or EXPERIENCERs. [1] model the information filtering structures of opinions and facts with a supervised approach to identify the hierarchical structure of perspective and speech expressions using syntactic dominance features, and to recursively determine local and global parent-child relations amongst such ex-

Table 5: *Multi-entity scoring results*

		ALL_POL		NON_NTR		Ranks (ALL_POL)				Errors (ALL_POL)		
Data set	Scoring	Acc	k	Acc	k	1	2	3	1+2	FATAL	GREEDY	LAZY
GS_SNTC	HUMAN	62.07	.45	89.03	.78					17.99	41.01	41.01
	COMPOS	52.20	.28	71.71	.45					38.66	38.13	**23.20**
	DIST	**56.44**	**.35**	79.32	.59	**56.44**	28.04	15.52	**84.48**	28.32	35.69	35.99
	SVM	50.04	.28	79.49	.58	50.04	30.64	19.31	80.69	**14.60**	**14.11**	71.28
	DIST+SVM	54.12	.33	**82.21**	**.64**	54.12	30.31	15.56	84.44	16.03	19.56	64.42
GS_PHR	HUMAN	61.63	.43	85.81	.70					18.38	40.81	40.81
	COMPOS	48.70	.24	65.56	.34					32.28	44.48	**23.23**
	DIST	51.42	**.27**	68.73	.40	51.42	27.51	21.07	78.93	27.41	39.68	32.91
	SVM	52.74	.25	**77.70**	**.52**	52.74	24.73	22.53	77.47	**12.42**	**20.70**	66.88
	DIST+SVM	**52.92**	**.27**	73.60	.48	**52.92**	26.08	21.00	**79.00**	18.52	28.71	52.77

pressions. However, only SOURCEs were targeted. A global Integer Linear Programming-driven constraint-based inference approach was used in [2] for joint extraction of sentiment expressions, SOURCEs, and their link relations using sequence tagging and relation classifiers with lexical, positional, and syntactic frame features. [7] extract HOLDERs and TOPICs using opinion verbs and adjectives, and FrameNet-driven semantic frame role labelling. In detecting HOLDERs, Maximum Entropy modelling with syntactic dependency features between sentiment expressions and candidate entities was used in [8]. [16], who highlight the insufficiency of automatic semantic role labelling in resolving SOURCEs and TARGETs, discuss the complexity involved in the task ranging from attribution, multiple SOURCEs and TARGETs, semantic scope, referents, discourse structure, inference, and TARGET relations, amongst others. The interrelation between sentiment roles and discourse structures is discussed further in [18] who propose transitive opinion frames for linking TOPICs. The role of co-reference resolution is discussed in [19] alongside a TOPIC annotation scheme that links opinions based on topical co-reference (cf. [6]).

6 Conclusion

This paper presents a principled, structural framework for modelling entity-level sentiment (sub)contexts, and in doing that, it sheds light on the role of (non-)compositional semantics in entity-level sentiment analysis. We demonstrated how compositional sentiment parsing lends itself naturally to multi-entity sentiment scoring with minimal modification. Initial results obtained from two scoring methods suggest that, despite the inherent complexity and subjectivity of the task, compositional sentiment parsing can generate sensible analyses that emulate human multi-entity sentiment judgements effectively.

References

[1] E. Breck and C. Cardie. Playing the telephone game: determining the hierarchical structure of perspective and speech expressions. In *Proceedings of COLING 2004*, pages 120–126, Geneva, Aug. 2004.

[2] Y. Choi, E. Breck, and C. Cardie. Joint extraction of entities and relations for opinion recognition. In *Proceedings of EMNLP 2006*, pages 431–439, Sydney, Jul. 2006.

[3] Y. Choi and C. Cardie. Learning with compositional semantics as structural inference for subsentential sentiment analysis. In *Proceedings of EMNLP 2008*, pages 793–801, Honolulu, Oct. 2008.

[4] X. Ding, B. Liu, and P. S. Yu. A holistic lexicon-based approach to opinion mining. In *Proceedings of the 1st ACM Intl. Conference on Web Search and Data Mining (WSDM 2008)*, pages 231–240, Palo Alto, Feb. 2008.

[5] M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of the ACM SIGKDD Intl. Conference on Knowledge Discovery & Data Mining (KDD 2004)*, pages 168–177, Seattle, Aug. 2004.

[6] J. S. Kessler and N. Nicolov. Targeting sentiment expressions through supervised ranking of linguistic configurations. In *Proceedings of the 3rd Intl. Conference on Weblogs and Social Media (ICWSM 2009)*, San Jose, May 2009.

[7] S.-M. Kim and E. Hovy. Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of the COLING/ACL 2006 Workshop on Sentiment and Subjectivity in Text*, pages 1–8, Sydney, Jul. 2006.

[8] S.-M. Kim and E. Hovy. Identifying and analyzing judgment opinions. In *Proceedings of HLT/NAACL-2006*, pages 200–207, New York, Jun. 2006.

[9] M. Klenner, S. Petrakis, and A. Fahrni. A tool for polarity classification of human affect from panel group texts. In *Proceedings of the Intl. Conference on Affective Computing and Intelligent Interaction (ACII 2009)*, Amsterdam, Sep. 2009.

[10] N. Kobayashi, K. Inui, and Y. Matsumoto. Extracting aspect-evaluation and aspect-of relations in opinion mining. In *Proceedings of EMNLP/CoNLL 2007*, pages 1065–1074, Prague, Jun. 2007.

[11] K. Moilanen and S. Pulman. Sentiment composition. In *Proceedings of RANLP 2007*, pages 378–382, Borovets, Sep. 2007.

[12] K. Nigam and M. Hurst. Towards a robust metric of opinion. In Y. Qu, J. Shanahan, and J. Wiebe, editors, *Computing Attitude and Affect in Text: Theory and Applications*, pages 265–280. Springer, 2006.

[13] A. Osherenko. Towards semantic affect sensing in sentences. In *Proceedings of the AISB 2008 Symposium on Affective Language in Human and Machine*, pages 41–44, Aberdeen, Apr. 2008.

[14] L. Polanyi and A. Zaenen. Contextual valence shifters. In Y. Qu, J. Shanahan, and J. Wiebe, editors, *Computing Attitude and Affect in Text: Theory and Applications*, pages 1–10. Springer, 2006.

[15] A.-M. Popescu and O. Etzioni. Extracting product features and opinions from reviews. In *Proceedings of HLT/EMNLP 2005*, pages 339–346, Vancouver, Oct. 2005.

[16] J. Ruppenhofer, S. Somasundaran, and J. Wiebe. Finding the sources and targets of subjective expressions. In *Proceedings of LREC 2008*, Marrakech, May 2008.

[17] M. A. M. Shaikh. *An Analytical Approach for Affect Sensing from Text*. PhD thesis, University of Tokyo, 2008.

[18] S. Somasundaran, J. Wiebe, and J. Ruppenhofer. Discourse level opinion interpretation. In *Proceedings of COLING 2008*, pages 801–808, Manchester, Aug. 2008.

[19] V. Stoyanov and C. Cardie. Annotating topics of opinion. In *Proceedings of LREC 2008*, Marrakech, May 2008.

[20] J. Wiebe, T. Wilson, and C. Cardie. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3):165–210, May 2005.