Generating Shifting Sentiment for a Conversational Agent

Simon Whitehead
University of Melbourne, Australia
srwhitehead@gmail.com

Lawrence Cavedon
RMIT University, Australia
lawrence.cavedon@rmit.edu.au

Abstract

We investigate techniques for generating alternative output sentences with varying sentiment, using (an approximation to) the Valentino method, based on SentiWordNet, of Guerini et al. We extend this method by filtering out unacceptable candidate sentences, using bigrams sourced from different corpora to determine whether lexical substitutions are appropriate in the given context. We also compare the generated candidates against human judgements of whether the desired sentiment shift has occurred: our results suggest limitations with the overall knowledge-based approach, and we propose potential directions for improvement.

1 Introduction

The design of more natural or *believable* conversational agents (Bates, 1994; Pelachaud and Bilvi, 2003) requires the need for such agents to communicate affectively, by the display of emotion or attitude towards objects, other agents, or states of affairs. More engaging or influential agents may seek to actually affect their conversational partner at a deeper level, for example, by influencing their emotional state (van der Sluis and Mellish, 2008). Previous work in this area has explored the use of gestures and facial expression (Caridakis et al., 2007) and rhythm and prosody of speech (Zovato et al., 2008) for expressing affect; however there has been little work on generation of affective language in dialogue.

Our general approach is inspired by (Fleischman and Hovy, 2002)'s work on generating different surface-level versions of utterance content, depending on an agent's appraisals towards objects, characters and events in its environment. While their approach is effective, it relies on manual creation of lexical alternatives, customized to the application domain. We are interested in approaches that will scale, and can be applied domain-independently.

While our ultimate aim is generation of language that relects emotional state, in this work we investigate the automatic generation of varying "sentiment" in output utterances; we focus on sentiment mainly due to the recent development of useful resources for this task. (Guerini et al., 2008)'s Valentino system is an approach to automatically generating candidate output utterances with different sentiment from an original; the authors suggest ECAs as a possible application scenario for their techniques. We explore this suggestion, implementing a *lexical substitution* (McCarthy and Navigli, 2007) approach to dialogue generation with sentiment, using the Valentino approach and associated resources. Lexical substitution approaches raise well-known challenges, and we investigate a number of techniques to address these in Section 4; for example, using bigrams and grammatical relations to determine which substitutions are acceptable based on their context in a sentence.[1]

Our techniques show improvement over naive lexical substitution; however, an evaluation with human subjects suggests that a deeper problem is that even "acceptable" candidate sentences generated by the method do not match human judgements with respect to sentiment shift: i.e., alternatives labeled as more positive (resp., negative) than the original by the system are often seen as a sentiment shift in the opposite direction by human judges (Section 5).

2 Background: Valentino

The *Valentino*[2] system (Guerini et al., 2008) is a tool developed from WordNet and SentiWordNet

[1] Guerini et al. suggest this as an area for further work.

[2] VALENced Text INOculator

Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, pages 89–97,
Los Angeles, California, June 2010. ©2010 Association for Computational Linguistics

designed to produce more positively or negatively slanted versions of text. Input to the system consists of a short sentence, and a *target valence* (between -1 and 1), which indicates the *desired* polarity and magnitude of sentiment in the modified output text. Valentino uses a number of strategies for adding, removing, or substituting certain words in order to alter the overall sentiment of the sentence. Table 1 shows examples of Valentino output for different target valences, with modifications in italics.

To perform the word-substitution, (Guerini et al., 2008) created a resource of *OVVTs*[3]: vectors of semantically related terms which may substitute for one another. The OVVTs were constructed using structural analysis of WordNet, and are divided into adjectives, nouns, and verbs. (Guerini et al., 2008) also constructed a separate resource of *Modifier OVVTs* which list adverbs that can be used to modify verbs. Modifier OVVTs were created using verbs extracted from certain *FrameNet*[4] categories, then recording which adverbs occur next to these verbs in the British National Corpus (BNC). Each term in the Valentino resource was assigned a *sentiment valence*, which corresponds to the SentiWordNet score of its parent WordNet synset. Table 2 shows part of an OVVT containing the noun 'man'.[5]

Term	POS	Sense	Valence
hunk	n	1	0.375
man	n	1	0
dude	n	1	-0.125
beau	n	2	-0.125

Table 2: (Abridged) example of an OVVT

To generate a modified sentence, (Guerini et al., 2008) apply the following strategies to each word[6] until the sentence valence (total of term valences) meets the target:

1. **Paraphrase:** Lemmas with only one sense are replaced by their WordNet gloss, which is scored for sentiment using the OVVTs;
2. **Use of most frequent senses:** The OVVTs are searched using only the most frequent senses;
3. **Adjective modification:** Adjectives are replaced with their stronger/weaker alternatives such that the target valence is not exceeded;
4. **Verb modification:** Verbs are modified by inserting, removing, or replacing intensifier or downtoner adverbs.

The final sentence is rendered as surface text by transforming each of the inserted lemmas back into the original morphology.

(Guerini et al., 2008) suggest their system's potential application to dialogue generation in an ECA, enabling emotional variation. However, they do not present an evaluation of Valentino's effectiveness. We expect that not all output utterances generated using their method will be sensible in the context of a believable ECA, for the following reasons:

Unconventional Word Usage: Upon inspection, we found the OVVTs often contain several words which are no longer conventionally used (e.g. "beau"). For an ECA to be believable, we hypothesise that such unpopular words should not be considered as potential candidates for substitution.

Incorrect Grammatical Context: The naive version of the Valentino method assumes that all words in an OVVT can be substituted for one another regardless of their context in the sentence (see Table 3); Guerini et al. propose this as an area for future work. We explore semi-informed solutions using bigrams and grammatical relations to eliminate syntactically incorrect substitutions.

... Williams was not *interested* (in) girls
... Williams was not *concerned* (with) girls
... Williams was not *fascinated* (by) girls

Table 3: Illustration of grammatical context issues

3 Implementation

We implemented a lexical substitution approach to varying valence, closely following the Valentino approach described in (Guerini et al., 2008). We did

[3] We assume OVVT stands for Ordered Vector of Valenced Terms; this is not explicit in (Guerini et al., 2008).

[4] http://framenet.icsi.berkeley.edu/

[5] All our examples and evaluations are using a version of the OVVTs made available by Marco Guerini on May 13, 2009.

[6] Actually, to the lemma of each word.

Valence	Sentence
n/a	Bob admitted that John is absolutely the best guy
1.0	Bob *wholeheartedly admitted* that John is *absolutely a superb hunk*
0.5	Bob *openly admitted* that John is *highly* the *redeemingest signor*
0.0	Bob *admitted* that John is *highly a well-behaved sir*
-0.5	Bob *sadly confessed* that John is *nearly a well-behaved beau*
-1.0	Bob *harshly confessed* that John is *pretty an acceptable eunuch*

Table 1: Example of Valentino sentiment shifting (Guerini et al., 2008)

not implement all the above strategies—in particular, we did not implement paraphrasing, adverb modification, or morphology synthesis; rather we focused on developing techniques that would address the lexical substitution issues described above.

As with Valentino, we calculate *sentence valence* by summing the valences of all terms in the sentence which are present in the OVVTs[7]. However, as a variation on Valentino, we aggregated sentence shift into five broad categories: "major positive shift"; "minor positive shift"; "no shift"; "minor negative shift"; "major negative shift".

Since most OVVTs contain only lemmas, we first performed *lemmatisation* using the *MorphAdorner*[8] package. To locate a term in the OVVTs, we first search for the original word morphology, then if no match is found we try using the lemma.

As with (Guerini et al., 2008), we included candidates from multiple senses of a matching word; however, rather than stopping at the third most frequent sense, we explored up to sense forty so as to increase the number of possible substitutions for terms.[9] We performed a very naive version of word sense disambiguation (WSD) (see below), but lack of WSD was an issue (discussed later).

Alternative sentences were generated by modifying at most a single word; this reduces the explosion in the number of alternatives, but the methods described could just as easily apply to alternatives constructed by varying multiple words.

The novel aspect of our implementation was the "candidate filtering" techniques: i.e. techniques for deciding whether to accept a candidate replacement term as substitute in a given sentence; this was specifically designed to address the issues above. In the next section, we describe filtering techniques using simple bigrams and grammatical relations, and evaluate the effectiveness of each.

4 Evaluation: Candidate Filtering

The data set we used for this evaluation consisted of 25 sentences, randomly extracted from the BNC.[10] The sentences were sourced from the BNC to avoid any bias which may have been introduced had the test sentences been created manually. We required that each test sentence satisfy the following conditions[11]:

1. The sentence must contain between 6 and 10 words (to reflect length of a typical dialogue utterance);

2. The sentence must contain at least one term which is found in the OVVTs (otherwise it would be pointless for evaluation purposes); the term may have any valence.[12]

Our second filtering technique requires information about the grammatical relations between terms in a sentence (illustrated in Figure 1). For this, we used a version of the BNC which was pre-processed with the RASP parser (Briscoe et al., 2006).

Our gold standard for candidate acceptability was created using the first author's judgements.[13] In or-

[7] Since we ignore adverbs, we do not include these when scoring a sentence.

[8] `http://morphadorner.northwestern.edu/`

[9] Increasing this further increased the number of alternatives but did not improve performance.

[10] The size of our test data set was capped at 25 due to the time required to create the gold standard (i.e., judging 1030 substitutions consistently).

[11] These constraints reduced our sample set from the ∼4.6 million sentences in the BNC to approx. 627,000 sentences.

[12] The sentence can theoretically be valence-shifted by substituting that term, regardless of the term's valence.

[13] With more time we would of course have preferred to use multiple annotators. However, the judgement task was simple

der to be judged as an ACCEPT by the annotator, a generated sentence needed to satisfy the following criteria (otherwise it was labelled REJECT):

1. **Semantic Equivalence:** The new sentence should convey reasonably equivalent semantics compared to the original: e.g., phrases such as 'young boy' and 'small boy' were considered acceptably close;[14]

2. **Grammatical Correctness:** The new sentence should not contain grammatical errors. For the gold standard, terms were *manually* converted into their original morphological form before annotation (e.g., if the lemma 'speak' replaced an instance of 'shouted', then it was converted to 'spoke').

4.1 Evaluation Methodology

To evaluate each candidate selection method, we performed the following procedure for each of our 25 test sentences:

1. Find all matching[15] terms and retrieve the valence score of each;

2. For each matching term:

   (a) Retrieve the corresponding list of alternative terms from the OVVTs;

   (b) Generate several different candidate sentences by substituting each alternative term into the original sentence;

   (c) Apply the chosen *candidate selection* technique to each generated sentence, and label each as ACCEPT or REJECT (for step 3);

3. Compare all system classifications to our gold standard (automatically), and mark each as either a true positive (TP), false positive (FP), true negative (TN), or false negative (FN).

We then used the TP, FP, TN and FN counts to compute the accuracy, precision, recall and F-score across all generated sentences. These metrics are used to compare the relative performance between each of our candidate selection methods.

We describe each of our techniques and the results; we present all the measurements in a single table (Table 5).[16]

4.2 Candidate filtering using bigrams

For each candidate sentence generated, we examined the bigrams including the newly substituted term. If both[17] bigrams appear in the BNC, we take this as an indication that the substitution is acceptable, and we accept the candidate sentence. Otherwise, the candidate is rejected. We pre-processed the BNC to extract 8,463,295 unique bigrams, formatted as `lemma/pos lemma/pos` pairs, where `lemma` is the lemmatised word, and `pos` is the WordNet POS. As a simple attempt to address word-sense disambiguation, we discriminated on POS[18] when extracting and matching these bigrams. For example, '`drive/n home/n`' and '`drive/v home/n`' would be considered separate bigrams, as the term 'drive' occurs with different POS in each. We chose to lemmatise all bigrams due to the relatively small size of the BNC. Also, we did not consider bigrams which are interrupted by sentence punctuation, as this indicates a phrase break.

We take this bigram approach as our baseline.[19] This simple technique has reasonable accuracy (0.752: see Table 5) but this is due largely to the high number of true negatives produced. The false negatives are mainly caused by the BNC's relatively limited bigram coverage.

To address this issue, we sourced our bigrams from the Google Web 1T Corpus, which covers approximately *one trillion* words of English text sourced from publicly accessible web pages. Compared with the BNC, it has much greater coverage, containing $\sim$314 million bigrams. However, Web 1T does not contain POS information, and due to its size we did not lemmatise the bigrams. Using a

enough for us to believe it to be reliable.

[14] A fairly liberal view of "semantic equivalence" was taken; for example, for our purposes we consider all sentences in Table 1 to be more-or-less semantically equivalent.

[15] A matching term is defined as a term which has a corresponding entry in the OVVTs.

[16] Note that had we performed *no filtering*, all TN's would become FP's and all FN's would befome TP's.

[17] For terms beginning/ending a sentence (or phrase surrounded by punctuation), we only examine one bigram.

[18] We differentiated only adjectives, nouns, verbs, and adverbs; all other POS were considered equivalent for the purposes of bigram extraction.

[19] A lower baseline would be to perform no filtering.

smaller corpus, these differences may reduce coverage and bigram matching accuracy. However we hypothesise that using the Web 1T corpus, such limitations should be outweighed by its sheer size.

From Table 5, we see a substantial increase in recall over our previous baseline, which supports our hypothesis that using a larger corpus would increase true positives and reduce false negatives. However, the increased coverage of the Web 1T corpus brings with it more opportunities for false positives, the number of which has increased dramatically from our baseline, causing a reduction in precision and accuracy. Despite this, due to increased recall, we achieved an improvement in overall F-score.

Due to its web-based nature, the Web 1T corpus will contain more errors than a corpus sourced from published print, such as the BNC. Bigrams which occur infrequently may be a source of noise. We hypothesized that a substitution is acceptable if its replacement bigrams occur in some reasonable proportion to the original bigrams. Hence, we experimented with *bigram frequency ratios*, where a candidate is accepted only if its ratio exceeds a given threshold The ratio is calculated as f_r/f_o, where f_r and f_o represent the replacement and original bigram frequencies, respectively. We repeated our Web 1T bigrams experiment for several ratio thresholds between 0 and 0.9, and measured the changes in accuracy, precision and recall. Our results showed that frequency ratio thresholding can reduce false positives, leading to slightly increased precision for certain ratios. However, true positives are also reduced, and we sacrifice significant recall for only minor gains in precision.

4.3 Filter using grammatical relations

Candidate selection using bigrams is a somewhat naïve approach, as it considers only the surface text without regard for the underlying *grammatical relations* (GRs) between terms. To illustrate, consider the example shown in Table 4.

Context	on their *lovely* family holidays
Term	lovely
Alt.s	handsome, picturesque, pretty, splendid, scenic, resplendent, ...

Table 4: Sample context & replacements for ‘lovely’

We observed that alternatives for ‘lovely’ such as ‘picturesque’ and ‘scenic’ were falsely rejected using BNC bigrams.[20] As bigrams, “picturesque family” and “scenic family” seem like unnatural ways of describing a family. However, in this context ‘lovely’ modifies ’holiday’, not ‘family’: this distinction is not picked up using simple bigrams. To address this limitation, we extended our bigram candidate selection technique to consider grammatical relations (GRs).

Our GR technique uses an input sentence in RASP format. We only change one term per sentence as before; however we first extract the term’s GRs from the RASP annotation. We convert each binary[21] GR into a *GR-bigram* using the original ordering of terms in the sentence. Figure 1 illustrates the GRs for our example sentence, and how such translate into GR-bigrams.

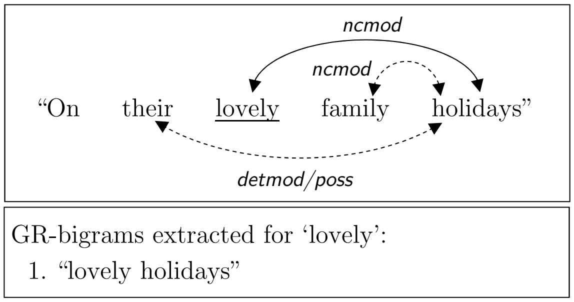

ncmod
ncmod
“On their lovely family holidays”
detmod/poss
GR-bigrams extracted for ‘lovely’:
1. “lovely holidays”

Figure 1: Grammatical relations and GR-bigrams

By converting GRs into bigrams, we can take advantage of Web 1T’s extensive coverage. However, due to our restrictions on GR types, it is possible to obtain zero GR-bigrams for some words in a sentence. This happens when the word has no modifier or comparative relations associated with it. For these words, we revert to our bigram selection technique.

Our results for candidate selection using GRs are again shown in Table 5. Surprisingly, this technique performs worse than using regular bigrams for all metrics when compared to our baseline. We suspect our GR selection technique performs no better than

[20] These candidates were accepted using the Web 1T corpus.

[21] We only examine binary comparative and modifier GR types, as RASP provides many other syntactic relations which we deemed not relevant to our task.

Web 1T bigrams simply due to the corpus' extensive coverage, which leads to a similar amount of false positives.

Selection Technique	BNC Bigrams	Web 1T Bigrams		Web 1T GRs	
True positives	22	**55**	**150%**	54	145%
False positives	**45**	155	244%	169	276%
True negatives	**288**	178	-38%	164	-43%
False negatives	57	**24**	**-58%**	25	-56%
Accuracy	**0.752**	0.566	-25%	0.529	-30%
Precision	**0.328**	0.262	-20%	0.242	-26%
Recall	0.278	**0.696**	**150%**	0.684	145%
F-score	0.301	**0.381**	**26%**	0.358	19%

Table 5: Collated results for all experiments

4.4 Error Analysis

To explain our experimental results, we first look at how the performance changes between our different versions relative to the baseline (i.e., BNC Bigrams): see Table 5. Note first that, while all methods increased the number of true positives and decreased false negatives, any performance gains were simply drowned out by the massive increases in false positives that occurred: this is the main cause of our low precision and recall. For the following discussion, we focus on the use of Web IT bigrams, which was the best performing filtering technique.

Since false positives are the most important source of error to avoid in an ECA, we focus on these. We examined the false positive instances and categorised each error into the following four groups. The distribution of errors into these categories is shown in Table 6.

Category	No. FP	% of all FP
Change in Meaning	76	49.03%
Incorrect WSD	42	27.10%
Phrase/Metaphor	31	20.00%
Grammatical	6	3.87%
Total	155	100%

Table 6: Distribution of classification errors

4.4.1 Change in meaning

A major limitation of the OVVT resource is that several of the alternative terms simply cause too much semantic change even when the correct sense of the original term is detected. For example, some alternatives for 'winner' are words such as 'sleeper', 'upsetter', and 'walloper'. In the context of the phrase "Cash prizes will be offered to the winners", we will almost always prefer the generic 'winner'.

We suspect this limitation arises due to the methods used to construct the OVVTs; in particular the use of the WordNet `hyponym` and `hypernym` relations. For example, the 'thing' category in WordNet encompasses a multitude of more specific terms, such as 'ornament', 'structure', 'surface', and 'installation'. These terms all made their way into the OVVT for 'thing', yet they are rarely appropriate substitutions for 'thing'. Conversely, we may not wish to replace any specific terms with the more generic 'thing' as this removes too much meaning.

As this kind of error accounted for almost half of our false positives, addressing this limitation may lead to significant gains in performance. This likely requires a more conservative approach to constructing the OVVTs themselves, e.g., by incorporating corpus-based information, as per (Guerini et al., 2008)'s approach to constructing the Modifier-OVVTs): the technique for mining appropriate verb-adverb pairings from the BNC could be generalised to include other POS types.

Related to the problem of semantic change is the idea of context-dependent semantics. For example, certain qualifiers have opposing effects depending on the appraisal of the subject: consider a "long term *illness*" compared to a "long term *vacation*". One possible solution to this problem is to modify the way valences are calculated to take into account which terms modify one another.

4.4.2 Incorrect word-sense disambiguation

The WSD approach used in our work adapted from (Guerini et al., 2008) is only a crude approximation to a complex problem; the WSD-related problems could at least be alleviated by incorporating a more sophisticated WSD approach into the pipeline. However, even if we could determine the correct sense of each word, we are still left with the limitation that the OVVTs are not exhaustive in their coverage, with several word senses missing.

4.4.3 Phrases and metaphors

Several false positives were caused by phrases such as "long term". Metaphors were a similar cause for error, e.g. "stepping stone". Phrase and metaphor detection should improve our technique's performance, especially since the OVVTs contain several phrases; however, these are known difficult challenges in themselves.

4.4.4 Grammatical errors

A grammatical error occurs when the alternative term is acceptable *semantically*, yet further syntactic modification to the sentence is needed to preserve correct grammar: see Table 3.

An extension of our bigram approach could be to use a larger window around replaced words to assess the suitability of a substitution. Recent work has shown this technique could be used to rank potential substitutions in order of acceptability (Hawker, 2007) and is worth considering as future work.

4.4.5 Limitations of bigrams and corpus coverage

In some cases, our bigram selection technique is ineffective when the term being changed is flanked by *stop words*. In a corpus of sufficient size and coverage, the majority of terms will occur next to stop words far more often than they occur next to other, less common terms. Hence, bigrams containing stop words were a common source of false positives.

This limitation could be addressed in future work by extending our grammatical relation technique to include *ternary* GRs, which provide relations for noun-verb phrases such as "solution to fitness" and "solution to health". Given these, we could accept or reject based on the presence of the accompanying *tri*grams in the Web 1T corpus. As described in (Hawker, 2007), use of an even larger window, such as 4-grams and 5-grams around replaced terms may also address this issue, however the size of the Web 1T corpus for larger N-grams presents serious processing challenges.[22]

5 Evaluation: Sentiment Shift

The technqiues described above attempt to create acceptable candidates to shift sentiment. However, this leaves open the question as to whether the technique has its desired effect: i.e. appropriately shifting sentiment. We designed an experiment which aims to measure correlation between human judgements of the sentiment shift in our generated candidates, and our system's representation of sentiment shift.

We presented subjects with an original sentence, along with *one* of the generated candidates. Our six subjects had no specialised knowledge of the task and were all native English speakers. Subjects were asked to judge the modified sentence for *change* in sentiment relative to the original according to the five shift categories described earlier (i.e., major/minor positive/negative/no shift). In order to avoid bias and to clarify the task, we explained that *sentiment* should be separated from changes in meaning, or the reader's opinions about the sentences. Instead, we urged subjects to ask themselves the question: "Is the author of the second sentence saying what they're saying in a more positive or more negative way, compared to the first sentence?"

The sentences used were extracted from the BNC at random, using the restrictions listed above. We extracted 250 sentences to be used as the originals, each of which was used as input to our sentiment shifting system. For each original sentence, we produced all possible candidates using our best performing candidate selection method, Web 1T Bigrams. We also limited our generation to changing one term per sentence, as to not produce a combinatorial explosion in the number of candidates generated. This produced approximately 3000 modified candidates, including several candidates with no sentiment shift.

Upon inspection, we found many generated candidates contained the types of errors described above. Hence, we manually extracted original and modified sentences until we had a total of 50 originals, and 100 shifted sentences. In selecting which sentences to keep, we chose ones which sounded the most natural, or had the least amount of semantic change from the original. Manual selection was performed in order to prevent introducing any bias into judgements when a subject is confronted with a grammatically incorrect or unnatural sentence. We also aimed for a fairly even distribution of the shifted

[22] (Hassan et al., 2007) describes a successful approach to lexical substitution that combines multiple knowledge sources.

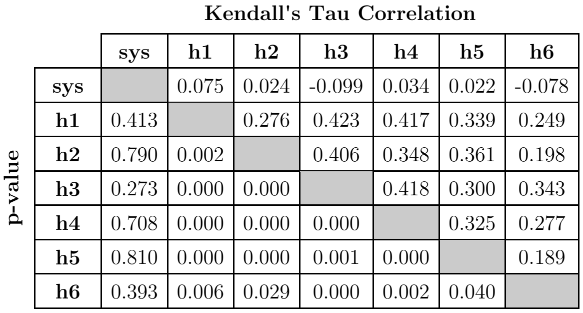
sentences into the five sentiment shift intervals.[23]

5.1 Results and analysis

We performed a pairwise Kendall's Tau rank correlation (Kendall and Gibbons, 1962), which compares each human's judgements with the system's sentiment shift, for all 100 generated sentences. Kendall's Tau measures the correlation between two distributions on a scale of -1 to 1, with 1 indicating total agreement; -1 indicating total disagreement; and 0 indicating no (or random) correlation.

We measured the correlation using the five sentiment shift intervals, and also using judgement *polarities*, i.e. whether a score is positive, negative or zero. We only report on polarity results as the finer-grained comparison showed similar results with slightly less correlation.

Our results are shown in Table 7; Kendall's Tau correlations are shown above the shaded diagonal, while the corresponding p-values for statistical significance are shown below the diagonal.



Kendall's Tau Correlation

p-value

	sys	h1	h2	h3	h4	h5	h6
sys		0.075	0.024	-0.099	0.034	0.022	-0.078
h1	0.413		0.276	0.423	0.417	0.339	0.249
h2	0.790	0.002		0.406	0.348	0.361	0.198
h3	0.273	0.000	0.000		0.418	0.300	0.343
h4	0.708	0.000	0.000	0.000		0.325	0.277
h5	0.810	0.000	0.000	0.001	0.000		0.189
h6	0.393	0.006	0.029	0.000	0.002	0.040	

Table 7: Kendall's Tau rank correlation between system (**sys**) and human (**hi**) judgement polarities

Although the correlation observed between inter-annotator judgements of polarity was fairly low, it is statistically significant in all cases using a confidence level of $p < 0.05$. While this indicates there was some agreement between human annotators, the relatively low correlation indicates that judging sentiment is a fairly subjective task. However, we saw no correlation between the human judgements and our system's representation of sentiment shift.

The poor correlation between human and system polarities can possibly be attributed to a number of reasons. (Guerini et al., 2008) mention that in SentiWordNet, several of the WordNet synsets are valenced incorrectly, with many having a valence of zero, which we also observed in the OVVT resource. Our survey results suggest that SentiWordNet in its current form is not ideally suited to the task of generating sentiment in text using the Valentino method.

SentiWordNet may be effective when classifying the sentiment of *large* texts; the valence scores can be considered to reflect the degree to which each word represents a sentiment "feature". However, it is somewhat unrealistic to assume that every term will have the same effect on sentiment in all contexts; assigning words a 'universal' sentiment score seems non-intuitive, and a finer-grained representation of sentiment is needed for short texts such as dialogue utterances.

In sentiment *generation*, when choosing a replacement term from a set of alternatives, we are more interested in each candidate's effect on sentiment, *relative* to the other candidates. While a resource of semantically clustered terms is needed for this task (such as the OVVTs), terms within each cluster need to be ranked for sentiment in a *localised* way, taking account of positivity or negativity relative to other terms in the cluster. Upon inspection of several OVVTs, this ranking is a straightforward task for a human to perform (if time-consuming).

However, the context of a substitution often determines its effects of sentiment. Hence, we argue that future work in sentiment generation using knowledge-based techniques should extend existing resources to encompass ranking of candidates in a *contextual* way, rather than ranking them statically out of context. For example, an MRE-style (Traum et al., 2003) approach could be used which goes beyond scoring the overall sentiment of an utterance, but considers how sentiment (or *attitude*) is directed towards agents, objects and events.

Acknowledgements: We thank Marco Guerini for kindly providing us with the OVVTs resource, Tim Baldwin for helpful suggestions for the evaluation in Section 5, and the referees for valuable feedback. Cavedon's contribution was partially supported by the Australian Research Council under Linkage Grant LP0882013.

[23] Note: the judgement of which sentiment-shift category a sentence-pair fell into was made by the system (and subjects); the manual intervention in the experiment design was to remove unacceptable sentence-pairs.

References

Joseph Bates. 1994. The Role of Emotion in Believable Agents. *Communications of the ACM*, 37(7):122–125.

Ted Briscoe, John Carroll, and Rebecca Watson. 2006. The Second Release of the RASP System. In *Proceedings of ACL*, pages 77–80, Sydney.

G. Caridakis, A. Raouzaiou, E. Bevacqua, M. Mancini, K. Karpouzis, L. Malatesta, and C. Pelachaud. 2007. Virtual Agent Multimodal Mimicry of Humans. *Language Resources and Evaluation*, 41(3):367–388.

Michael Fleischman and Eduard Hovy. 2002. Towards Emotional Variation in Speech-Based Natural Language Generation. In *Proceedings of the Second International Natural Language Generation Conference*, pages 57–64, New York.

Marco Guerini, Carlo Strapparava, and Oliviero Stock. 2008. Valentino: A Tool for Valence Shifting of Natural Language Texts. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, Marrakech.

Samer Hassan, Andras Csomai, Carmen Banea, Ravi Sinha, and Rada Mihalcea. 2007. Unt: Subfinder: combining knowledge sources for automatic lexical substitution. In *Proc. Fourth Int. Workshop on Semantic Evaluations (SemEval 2007)*, pages 410–413, Prague.

Tobias Hawker. 2007. USYD: WSD and Lexical Substitution Using the Web1T Corpus. In *Proc. 4th Int. Workshop on Semantic Evaluations (SemEval 2007)*, pages 446–453, Prague.

M.G. Kendall and J.D. Gibbons. 1962. *Rank Correlation Methods*. Griffin London.

Diana McCarthy and Roberto Navigli. 2007. SemEval-2007 task 10: English lexical substitution task. In *Proc. Fourth Int. Workshop on Semantic Evaluations (SemEval 2007)*, pages 48–53, Prague.

Catherine Pelachaud and Massimo Bilvi. 2003. Computational Model of Believable Conversational Agents. In *Communication in Multiagent Systems*, volume 2650 of *Lecture Notes in Computer Science*, pages 300–317. Springer.

David Traum, Michael Fleischman, and Eduard Hovy. 2003. NL Generation for Virtual Humans in a Complex Social Environment. In *In Proceedings of the AAAI Spring Symposium on Natural Language Generation in Spoken and Written Dialogue*, pages 151–158, Palo Alto.

Ielka van der Sluis and Chris Mellish. 2008. Towards Affective Natural Language Generation: Empirical Investigations. In *Proceedings of the Symposium on Affective Language in Human and Machine, AISB*, pages 9–16, Aberdeen.

E. Zovato, F. Tini Brunozzi, and M. Danieli. 2008. Interplay between pragmatic and acoustic level to embody expressive cues in a Text to Speech system. In *Proceedings of the Symposium on Affective Language in Human and Machine, AISB*, pages 88–91, Aberdeen.