Valence Shifting: Is It A Valid Task?

Mary Gardiner
Centre for Language Technology
Macquarie University
mary@mary.gardiner.id.au

Mark Dras
Centre for Language Technology
Macquarie University
mark.dras@mq.edu.au

Abstract

This paper looks at the problem of valence shifting, rewriting a text to preserve much of its meaning but alter its sentiment characteristics. There has only been a small amount of previous work on the task, which appears to be more difficult than researchers anticipated, not least in agreement between human judges regarding whether a text had indeed had its valence shifted in the intended direction. We therefore take a simpler version of the task, and show that sentiment-based lexical paraphrases do consistently change the sentiment for readers. We then also show that the Kullback-Leibler divergence makes a useful preliminary measure of valence that corresponds to human judgements.

1 Introduction

This paper looks at the problem of VALENCE SHIFTING, rewriting a text to preserve much of its meaning but alter its sentiment characteristics. For example, starting with the sentence *If we have to have another parody of a post-apocalyptic America, does it have to be this bad?*, we could make it more negative by changing *bad* to *abominable*. Guerini et al. (2008) say about valence shifting that

> it would be conceivable to exploit NLP techniques to slant original writings toward specific biased orientation, keeping as much as possible the same meaning ... as an element of a persuasive system. For instance a strategic planner may decide to intervene on a draft text with the goal of "coloring" it emotionally.

There is only a relatively small amount of work on this topic, which we review in Section 2. From this work, the task appears more difficult than researchers originally anticipated, with many factors making assessment difficult, not least the requirement to be successful at a number of different NLG tasks, such as producing grammatical output, in order to properly evaluate success. One of the fundamental difficulties is that it is difficult to know whether a particular approach has been successful: researchers have typically had some trouble with inter-judge agreement when evaluating whether their approach has altered the sentiment of a text. This casts doubt on valence shifting as a well-defined task, although intuitively it should be, given that writing affective text has a very long history, computational treatment of sentiment-infused text has recently been quite effective.

This encourages us to start with a simpler task, and show that this simpler version of valence shifting achieves agreement between judges on the direction of the change. Consequently, in this paper we limit ourselves to exploring valence shifting by lexical substitution rather than exploring richer paraphrasing techniques, and testing this on manually constructed sentences. We then explore two questions:

1. Is it in fact true that altering a single lexical item in a sentence noticeably changes its sentiment for readers?
2. Is there a quantitative measure of relative lexical valence within near-synonym sets that corresponds with human-detectable differences in valence?

We investigate these questions for negative words by means of a human experiment, presenting

Mary Gardiner and Mark Dras. 2012. Valence Shifting: Is It A Valid Task? In *Proceedings of Australasian Language Technology Association Workshop*, pages 42−51.

readers with sentences with a negative lexical item replaced by a different lexical item, having them evaluate the comparative negativity of the two sentences. We then investigate the correspondence of the human evaluations to certain metrics based on the similarity of the distribution of sentiment words to the distribution of sentiment in a corpus as a whole.

2 Related work

2.1 Valence-shifting existing text

Existing approaches to valence shifting most often draw upon lexical knowledge bases of some kind, whether custom-designed for the task or adapted to it. Existing results do not yet suggest a definitively successful approach to the task.

Inkpen et al. (2006) used several lexical knowledge bases, primarily the near-synonym usage guide *Choose the Right Word* (CtRW) (Hayakawa, 1994) and the General Inquirer word lists (Stone et al., 1966) to compile information about attitudinal words in order to shift the valence of text in a particular direction which they referred to as making "more-negative" or "more-positive". They estimated the original valence of the text simply by summing over individual words it contained, and modified it by changing near synonyms in it allowing for certain other constraints, notably collocational ones. Only a very small evaluation was performed involving three paragraphs of changed text, the results of which suggested that agreement between human judges on this task might not be high. They generated more-positive and more-negative versions of paragraphs from the British National corpus and performed a test asking human judges to compare the two paragraphs, with the result that the system's more-positive paragraphs were agreed to be so three times out of nine tests (with a further four found to be equal in positivity), and the more-negative paragraphs found to be so only twice in nine tests (with a further three found to be equal).

The VALENTINO tool (Guerini et al., 2008; Guerini et al., 2011) is designed as a pluggable component of a natural language generation system which provides valence shifting. In its initial implementation it employs three strategies, based on strategies employed by human subjects: modifying single wordings; paraphrasing, and deleting or inserting sentiment charged modifiers. VALENTINO's strategies are based on part-of-speech matching and are fairly simple, but the authors are convinced by its performance. VALENTINO relies on a knowledge base of Ordered Vectors of Valenced Terms (OVVTs), with graduated sentiment within an OVVT. Substitutions in the desired direction are then made from the OVVTs, together with other strategies such as inserting or removing modifiers. Example output given input of (1a) is shown in the more positive (1b) and the less positive (1c):

(1) a. We ate *a very good dish.*
    b. We ate *an incredibly delicious dish.*
    c. We ate *a good dish.*

Guerini et al. (2008) are presenting preliminary results and appear to be relying on inspection for evaluation: certainly figures for the findings of external human judges are not supplied. In addition, some examples of output they supply have poor fluency:

(2) a. * Bob *openly admitted* that John is *highly* the *redeemingest signor.*
    b. * Bob *admitted* that John is *highly a well-behaved sir.*

Whitehead and Cavedon (2010) reimplement the lexical substitution, as opposed to paraphrasing, ideas in the VALENTINO implementation, noting and attempting to address two problems with it: the use of unconventional or rare words (*beau*), and the use of grammatically incorrect substitutions.

Even when introducing grammatical relation-based and several bigram-based measures of acceptability, they found that a large number of unacceptable sentences were generated. Categories of remaining error they discuss are: large shifts in meaning (for example by substituting *sleeper* for *winner*, accounting for 49% of identified errors); incorrect word sense disambiguation (accounting for 27% of identified errors); incorrect substitution into phrases or metaphors (such as *long term* and *stepping stone*, accounting for 20% of identified errors); and grammatical errors (such as those shown in (3a) and (3b), accounting for 4% of identified errors).

(3) a. Williams was not *interested* (in) girls.
    b. Williams was not *fascinated* (by) girls.

Whitehead and Cavedon (2010) also found that their system did not perform well when evaluated. Human judges had low, although significant, agreement with each other about the sentiment of a sentence but not significant agreement with their system's output: that is, they did not agree if sentiment shifted in the intended way.

3 Human evaluation of valence shifting

We first describe the construction of our test data, followed by the process for eliciting human judgements on the test data.

3.1 Test data

3.1.1 Selection of negativity word pairs

Quite a number of lexical resources related to sentiment have been developed, and it may seem likely that there would be an appropriate one for choosing pairs of near-synonyms where one is more negative and the other less. However, none are really suitable.

- Several resources based on WordNet contain synsets annotated with sentiment information in some fashion: these include SentiWordNet (Esuli and Sebastiani, 2006), MicroWNOP (Cerini et al., 2007) and WordNet Affect (Strapparava and Valitutti, 2004), and a dataset of subjectivity- and polarity-annotated WordNet senses by Su and Markert (2008). Individual words within a synset are not, however, given individual scores, which is what we need.
- The General Inquirer word list (Stone et al., 1966), which contains unscored words in certain categories including positive (1915 words) and negative (2291 words), does not group words into sets of near-synonyms.
- The subjectivity lexicon that is part of the MPQA Opinion Corpus does assign terms to categories, in this case positive, negative, both or neutral, but does not score the strength of their affective meaning, although this corpus does rate their effectiveness as a cue for subjectivity analysis (Wiebe et al., 2005; Wilson et al., 2005).

The closest work to that described here is that of Mohammad and Turney (2010) and Mohammad and Turney (forthcoming), who describe in detail the creation of EmoLex, a large polarity lexicon, using Mechanical Turk. Mohammad and Turney (forthcoming), rather than asking annotators to evaluate words in context as we are proposing here, instead ask them directly for their analysis of the word, first using a synonym-finding task in order to give the worker the correct word sense to evaluate. Part of a sample annotation question given by Mohammad and Turney (forthcoming) is given in Table 1. The word source used is the *Macquarie Thesaurus* (Bernard, 1986).

Our work differs from that of Mohammad and Turney (forthcoming) in that we rely on substitution evaluations, that is, having human judges rate specific contexts rather than supply their intuitions about the meaning of a word. Callison-Burch (2007) argued for this evaluation of paraphrases, that the most natural way is through substitution, and evaluate both meaning and grammaticality.

In our case, we are attempting to assess the effectiveness of valence-shifting, and we cannot presuppose that intuitions by the raters along the lines of feeling that the meaning of a word is more negative than that of another word translates into perceiving the desired effect when a word is used in context.

We therefore turn to hand-crafted data to test our hypotheses: words chosen so as to be noticeably negative, with a neutral or slightly negative near synonym. We chose 20 such word pairs, shown in Table 2. The more negative word of the pair is from the sentiment lists developed by Nielsen (2011),[1] typically rated about 3 for negativity on his scale (where 5 is reserved for obscenities) and the less negative chosen by us.

3.1.2 Selection of sentences

Our corpus for sentence selection is the SCALE dataset v1.0 movie review data set (SCALE 1.0) (Pang and Lee, 2005), a set of 5000 short movie reviews by four authors on the World Wide Web, and widely used in sentiment classification tasks. Each movie review is accompanied by both a three and four degree sentiment rating (that is, a rating on a scale of 0 to 2, and on a scale of 0 to 3) together with original rating assigned by the author to their own review on a scale of 0 to 10.

[1] Available from `http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010`

Question	Possible answers
Which word is closest in meaning (most related) to startle?	{*automobile*, *shake*, *honesty*, *entertain*}
How positive (good, praising) is the word *startle*?	*startle* is {not, weakly, moderately, strongly} positive
How negative (bad, criticizing) is the word *startle*?	*startle* is {not, weakly, moderately, strongly} negative
How much is *startle* associated with the emotion {joy,sadness,...}?	*startle* is {not, weakly, moderately, strongly} associated with {joy,sadness,...}

Table 1: Sample annotation question posed to Mechanical Turk workers by Mohammad and Turney (forthcoming).

We selected two sentences for each word pair from the SCALE 1.0 corpus. Sentences were initially selected by a random number generator: each sentence originally contained the more negative word. Since we are constructing an idealised system here, evaluating the possibility of valence shifting by changing a single word, we manually eliminated sentences where the part of speech didn't match the intended part of speech of the word pair, where the word was part of a proper name (usually a movie title) and where the fluency of the resulting sentence otherwise appeared terribly bad to us. Where a sentence was rejected another sentence was randomly chosen to take its place until each word pair had two accepted sentences for a total of 40 sentences. We then made changes to capitalisation where necessary for clarity (for example, capitalising movie titles, as the corpus is normalised to lower case).

Since each subject is being presented with multiple sentences (40 in this experiment), rather than coming to the task untrained, it is possible that there are ordering effects between sentences, in which a subject's answers to previous questions influence their answers to following questions. Therefore we used a Latin square design to ensure that the order of presentation was not the same across subjects, but rather varied in a systematic way to eliminate the possibility of multiple subjects seeing questions in the same order. In addition, the square is balanced, so that there is no cyclical ordering effect (i.e. if one row of a Latin square is A-B-C and the next B-C-A, there is still an undesirable effect where C is tending to follow B). The presentation word order to subjects was also randomised at the time of generating each subject's questions.

3.2 Elicitation of judgements

Having constructed the set of test sentences (Section 3.1), we ask human subjects to analyse the sentences on two axes: ACCEPTABILITY and NEGATIVITY. This is loosely equivalent to the FLUENCY and FIDELITY axes that are used to evaluate machine translation (Jurafsky and Martin, 2009). As in the case of machine translation, a valence-shifted sentence needs to be fluent, that is to be a sentence that is acceptable in its grammar, semantics and so on, to listeners or readers. While some notion of fidelity to the original is also important in valence shifting, it is rather difficult to capture without knowing the intent of the valence shifting, since unlike in translation a part of the meaning is being deliberately altered. We therefore confine ourselves in this work to confirming that the valence shifting did in fact take place, by asking subjects to rate sentences.

In order to obtain a clear answer, we specifically evaluate valence shifting with sentences as close to ideal as possible, choosing words we strongly believe to have large valence differences, and manually selecting sentences where the subjects' assessment of the valence of these words is unlikely to be led astray by very poor substitutions such as replacing part of a proper name. (For example, consider the band name *Panic! at the Disco*: asking whether an otherwise identical sentence about a band named *Concern! at the Disco* is less negative is unlikely to get a good evaluation of lexical valence shifting.) We then ask human subjects to evaluate these pairs of sentences for their relative fluency and negativity.

Mechanical Turk Our subjects were recruited through Amazon Mechanical Turk.[2] Mechanical Turk is a web service providing cheap de-

[2] http://www.mturk.com/

centralised work units called Human Intelligence Tasks (HITs), which have been used by computational linguistics research for experimentation. Snow et al. (2008) cite a number of studies at that time which used Mechanical Turk as an annotation tool, including several which used Mechanical Turk rather than expert annotators to produce a gold standard annotation to evaluate their systems.

Callison-Burch and Dredze (2010) provide guidelines in appropriate design of tasks for Mechanical Turk, which we broadly follow. We ameliorate potential risks of using Mechanical Turk by confining ourselves to asking workers for numerical ratings of sentences, rather than any more complex tasks, well within the type of tasks which Snow et al. (2008) reported success with; and like Molla and Santiago-Martinez (2011), giving all subjects two elimination questions in which the sentences within each pair were identical, that is, in which there was no lexical substitution. These, being identical, should receive identical scores—we also explicitly pointed this out in the instructions—and therefore we could easily eliminate workers who did not read the instructions from the pool.

Eliciting subjects' responses We considered both categorical responses (e.g. Is sentence variant A more or less negative than sentence variant B, or are A and B equally negative?) and Magnitude Estimation (ME). Categorical responses of the sort exemplified ignore magnitude, and are prone to "can't decide" option choices.

ME is a technique proposed by Bard et al. (1996) for adapting to grammaticality judgements. In this experimental modality, subjects are asked evaluate stimuli based not on a fixed rating scale, but on an arbitrary rating scale in comparison with an initial stimulus. For example, subjects might initially be asked to judge the acceptability of *The cat by chased the dog*. Assuming that the subject gives this an acceptability score of N, they will be asked to assign a multiplicative score to other sentences, that is, $2N$ to a sentence that is twice as acceptable and $\frac{N}{2}$ to one half as acceptable.

This same experimental modality was used by Lapata (2001) in which subjects evaluated the acceptability of paraphrases of adjectival phrases, for example, considering the acceptability of each of (4b) and (4c) as paraphrases of (4a):

(4) a. a *difficult* customer
b. a customer that is *difficult* to *satisfy*
c. a customer that is *difficult* to *drive*

In a standard design and analysis of a ME experiment (Marks, 1974), all the stimuli given to the subjects have known relationships (for example, in the original psychophysics context, that the power level for one heat stimulus was half that of another stimulus), and the experimenter is careful to provide subjects with stimuli ranging over the known spectrum of strength under investigation. In our case, we do not have a single spectrum of stimuli such as a heat source varying in power, or even the varying degrees of fluency given by Bard et al. (1996) or the hypothesised three levels of paraphrase acceptability (low, medium, high) that Lapata (2001) is testing that her subjects can detect. Instead, we have distinct sets of stimuli, each a pair of words, in which we hypothesise a reliable detectable difference within the pair of words, but not between a member of one pair and a member of any other pair. Thus, asking subjects to rate stimuli across the pairs of words on the same scale, as ME requires, is not the correct experimental design for our task.

We therefore use an 11 point (0 to 10) rating scale. This allows subjects to rate two sentences as identical if they really perceive the sentences to be so, while allowing fairly subtle differences to be captured. This is similar to the assessment of machine translation performance used by NIST. For our fluency guidelines, we essentially use the ones given as NIST guidelines (Linguistic Data Consortium, 2005); we also model our negativity guidelines on these.

For each translation of each segment of each selected story, judges make the fluency judgement before the adequacy judgement. We provide similar questions to NIST, although with more context in the actual instructions. The precise wording of one of our questions is shown in Figure 1.

3.3 Results

3.3.1 Number of participants

A total of 48 workers did the experiment. 8 were excluded from the analysis, for these reasons:

1. 6 workers failed to rate the identical sentence pairs in the elimination questions described in Section 3.2 identically, contrary to explicit

Acceptability and negativity: concern/panic
Evaluate these two sentences for acceptability and negativity:

- Sentence 1: As they do throughout the film the acting of CONCERN and fear by Gibson and Russo is genuine and touching.
- Sentence 2: As they do throughout the film the acting of PANIC and fear by Gibson and Russo is genuine and touching.

Acceptability: first sentence of concern/panic pair
Give sentence 1 immediately above a score from 0 to 10 for its acceptability, where higher scores are more acceptable. The primary criterion for acceptability is reading like fluent English written by a native speaker.
Acceptability: second sentence of concern/panic pair
Give sentence 2 immediately above a score from 0 to 10 inclusive for its acceptability, where higher scores are more acceptable.
Negativity: first sentence of concern/panic pair
Give sentence 1 immediately above a score from 0 to 10 inclusive its negativity, where higher scores are more negative.
Negativity: second sentence of concern/panic pair
Give sentence 2 immediately above a score from 0 to 10 inclusive its negativity, where higher scores are more negative.

Figure 1: One of the acceptability and negativity questions posed to Mechanical Turk workers.

instructions.

2. 1 worker confined themselves only to the numbers 5 and 10 in their ratings.
3. 1 worker awarded every sentence 10 for both acceptability and negativity.

Each of the 8 Latin square rows were re-submitted to Mechanical Turk for another worker to complete.[3]

3.3.2 Analysing scaled responses

We consider two hypotheses:

1. that subjects will perceive a difference in *acceptability* between the original sentence and that containing a hypothesised less negative near synonym; and
2. that subjects will perceive a difference in *negativity* between the original sentence and that containing a hypothesised less negative near synonym.

We thus require hypothesis testing in order to determine if the means of the scores of the original sentences and those containing hypothesised less negative near synonyms differ significantly. In this situation, we can use a single-factor within-subject analysis of variance (ANOVA), also known as a single-factor repeated-measures ANOVA, which allows us to account for the fact that subjects are not being exposed to a single experimental condition each, but are exposed to all the experimental conditions. In this experiment we do not have any between-subjects factors—known differences between the subjects (such as gender, age, and so on)—which we wish to explore. A within-subjects ANOVA accounts for the lesser variance that can be expected by the subject remaining identical over repeated measurements, and thus has more sensitivity than an ANOVA without repeated measures (Keppel and Wickens, 2004). Our use of an ANOVA is similar to that of Lapata (2001), although we have only one factor. Specifically, we will test whether the *mean* scores of the more negative sample are higher than the less negative sample.

Acceptability results The mean acceptability rating of sentences containing the More Negative words from Table 2 was 6.61. The mean acceptability rating of sentences containing the Less Negative words was 6.41. An ANOVA does not find this difference to be statistically significant. ($F(1, 39) = 1.5975, p = 0.2138$). This is what we would expect: we manually selected sentences whose less negative versions were ac-

[3] In addition, one worker returned a single score of 610 for the negativity of one of the Less Negative sentences: we assume this was a data entry error and the worker intended either 6 or 10 as the value. In our analysis we set this value to 10, since it is the worse (i.e. most conservative) assumption for our hypothesis that sentences containing Less Negative words will have a lower negativity score than those containing More Negative words.

ceptable to us.

Negativity results The mean negativity rating of sentences containing the MORE NEGATIVE words from Table 2 was 6.11. The mean negativity rating of sentences containing the LESS NEGATIVE words was 4.82. An ANOVA finds this difference to be highly statistically significant. ($F(1, 39) = 29.324, p = 3.365 \times 10^{-6}$). In Table 2 we see that the effect is not only statistically significant overall, but very consistent: sentences in the LESS NEGATIVE group *always* have a lower mean rating than their pair in the MORE NEGATIVE group.

4 Predicting the raters' decisions

We now investigate to what extent we can predict the correct choice of near synonym so as to achieve the correct level of negativity in output. In the preceding section our data suggests that this can be accomplished with lexical substitution. However, this leaves the problem of determining the negativity of words automatically, rather than relying on hand-crafted data.

4.1 Measures of distribution

Our intuition is that words that make text more negative will tend to disproportionately be found in more negative documents, likewise words that make text less negative will tend to be found in less negative documents.

In order to quantify this, consider this as a problem of distribution. Among a set of affective documents divided into sentiment-score categories such as SCALE 1.0 (see Section 3.1), there is a certain, not necessarily even, distribution of words: for example, a corpus might be 15% negative, 40% neutral and 45% positive by total word count. However, our intuition leads us to hypothesise that the distribution of occurrences of the word *terrible*, say, might be shifted towards negative documents, with some larger percentage occurring in negative documents.

We then might further intuit that words could be compared by their relative difference from the standard distribution: a larger difference from the distribution implies a stronger skew towards some particular affective value, compared to word frequencies as a whole. (However, it should be noted that this skew could have any direction, including a word being found disproportionately among the neutral or mid-range sentiment documents.)

We thus consider two measures of differences of distribution, Information Gain (IG) and Kullback-Leibler divergence (KL). We calculate the value of our distribution measure for each MORE NEGATIVE and LESS NEGATIVE word pair, and subtract the former from the latter. If each word in the pair is distributed across sentiment categories in the same way, the difference will be zero; if the measure corresponds in some way to the human view of the word pair elements, the difference will be non-zero and have a consistent sign.

Information gain The IG $G(Y|X)$ associated with a distribution Y given the distribution X is the number of bits saving in transmitting information from Y if X is known. A high IG value thus suggests a strong predictive relationship between X and Y. We use the formulation of Yang and Pedersen (1997), who found it one of the more effective metrics for feature selection for text classification:

$$\begin{aligned} IG(r) = \quad & -\textstyle\sum_{i=1}^{m} \Pr(c_i) \log \Pr(c_i) \\ & + \Pr(r) \textstyle\sum_{i=1}^{m} \Pr(c_i|r) \log \Pr(c_i|r) \\ & + \Pr(\bar{r}) \textstyle\sum_{i=1}^{m} \Pr(c_i|\bar{r}) \log \Pr(c_i|\bar{r}) \quad (1) \end{aligned}$$

where $P_r(c_i)$ is the relative probability of category c_i, $P_r(c_i|t)$ the relative probability of c_i given term t and $P_r(c_i|\bar{t})$ the relative probability of c_i when term t is absent.

Kullback-Leibler Cover and Thomas (1991) describe the KL divergence (Kullback and Leibler, 1951) as a measure of "the inefficiency of assuming that the distribution is q when the true distribution is p". Weeds (2003) gives the formula for KL as:

$$D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \qquad (2)$$

Weeds (2003) evaluated measures similar to KL for their usefulness in the distributional similarity task of finding words that share similar contexts. Our task is not an exact parallel: we seek the relative skewness of words.

4.2 Results

Information Gain The results of the IG metric given in (1) on the test data are shown in the second column from the right in Table 2. No pattern

More Negative / Less Negative	MR	ΔMR	ΔIG $\times 10^{-3}$	ΔKL
ignored / overlooked	5.7/5.0	0.7	-1.48	0.12
cowardly / cautious	6.1/4.9	1.2	0.06	-0.04
toothless / ineffective	6.1/5.1	1.0	0.35	-0.39
stubborn / persistent	5.3/4.3	1.0	0.19	-0.31
frightening / concerning	6.2/5.5	0.7	-0.02	0.10
assassination / death	6.2/6.0	0.2	-0.99	-0.04
fad / trend	5.5/3.5	2.0	-0.09	0.00
idiotic / misguided	6.3/5.6	0.7	2.25	-0.27
war / conflict	6.5/5.4	1.1	2.03	-0.01
accusation / claim	6.3/4.5	1.8	0.35	-0.23
heartbreaking / upsetting	5.8/5.7	0.1	0.97	0.22
conspiracy / arrangement	5.6/4.1	1.5	-0.21	-0.02
dread / anticipate	6.6/3.9	2.7	1.58	-0.02
threat / warning	6.6/5.1	1.6	0.46	-0.10
despair / concern	6.2/4.5	1.7	0.21	-0.03
aggravating / irritating	6.2/5.7	0.5	-2.00	-0.09
scandal / event	6.9/3.8	3.1	-0.29	-0.09
panic / concern	6.5/4.5	2.0	0.56	-0.27
tragedy / incident	5.9/4.6	1.3	6.02	-0.08
worry / concern	5.3/4.5	0.7	0.31	-0.02

Table 2: More Negative / Less Negative word pairs; mean negativity ratings (MR); difference in mean negativity ratings (ΔMR); difference in Information Gain (ΔIG $\times 10^{-3}$); difference in Kullback-Leibler score (ΔKL)

in the data is immediately obvious, and in particular the ordering of More Negative and Less Negative is not maintained well by the metric.

Kullback-Leibler The results of the KL metric given in (2) on the test data are shown in the rightmost column of Table 2. Here we see a much stronger pattern, that the word from More Negative tends to have a lesser KL value than the word from Less Negative (16 out of 20 word pairs).

Preliminary indications are thus that the KL may be a more useful metric for predicting the raters' scores most accurately, and thus perhaps for predicting negativity in usage more generally.

5 Conclusion

In this paper, we have shown that lexical substitution, as we hoped, can achieve valence shifting on its own, as judged by human raters with a substitution task. In addition, we have shown that at least one measure of the distribution of a word in a corpus, the KL divergence, is a potentially promising feature for modelling the ability of a lexical substitution to achieve a valence shift. Valence shifting then, at least in this simplified form, would appear to be a well-founded task. However, successfully implementing a fuller version of valence shifting would face several challenges. A significant one is that no existing lexical resources are suitable as is. The use of the KL metric as a way of automatically scoring elements of near-synonym sets is preliminary, and would likely need further metrics, perhaps combined in a machine learner, to be able to accurately predict human judges' scores of negativity.

Acknowledgements

We acknowledge the support of ARC Discovery grant DP0558852.

References

Ellen Gurman Bard, Dan Robertson, and Antonella Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language*, 72(1):32–68, March.

John R. L. Bernard, editor. 1986. *The Macquarie Thesaurus*. Macquarie Library, Sydney, Australia.

2010. Los Angeles, California.

Chris Callison-Burch and Mark Dredze. 2010. Creating speech and language data with Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 1–12, Los Angeles, June. Association for Computational Linguistics.

Chris Callison-Burch. 2007. *Paraphrasing and Translation*. Ph.D. thesis, University of Edinburgh.

Sabrina Cerini, Valentina Compagnoni, Alice Demontis, Maicol Formentelli, and Caterina Gandi. 2007. Micro-WNOp: A gold standard for the evaluation of automatically compiled lexical resources for opinion mining. In *Language resources and linguistic theory: Typology, second language acquisition, English linguistics*. Franco Angeli, Milan, Italy.

Thomas M. Cover and Joy A. Thomas. 1991. *Elements of Information Theory*. Wiley, New York, USA.

Andrea Esuli and Fabrizio Sebastiani. 2006. SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006)*, pages 417–422, Genova, Italy.

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 (LREC 2008)*, Marrakesh, Morocco, May.

Marco Guerini, Carlo Strapparava, and Oliviero Stock. 2011. Slanting existing text with Valentino. In *Proceedings of the 16th International Conference on Intelligent User Interfaces (IUI '11)*, pages 439–440, Palo Alto, CA, USA, February.

Samuel I. Hayakawa, editor. 1994. *Choose the Right Word*. Harper Collins Publishers, 2nd edition. revised by Eugene Ehrlich.

Diana Zaiu Inkpen, Ol'ga Feiguina, and Graeme Hirst. 2006. Generating more-positive or more-negative text. In James G. Shanahan, Yan Qu, and Janyce Wiebe, editors, *Computing Attitude and Affect in Text (Selected papers from the Proceedings of the Workshop on Attitude and Affect in Text, AAAI 2004 Spring Symposium)*, pages 187–196. Springer, Dordrecht, The Netherlands.

Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*. Prentice-Hall, 2nd edition.

Geoffrey Keppel and Thomas D. Wickens. 2004. *Design and Analysis: A Researcher's Handbook*. Pearson Prentice Hall, Upper Saddle River, New Jersey, USA, fourth edition.

Solomon Kullback and Richard A. Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, 22:79–86.

Maria Lapata. 2001. A corpus-based account of regular polysemy: The case of context-sensitive adjectives. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 63–70, Pittsburgh, PA.

Linguistic Data Consortium. 2005. Linguistic data annotation specification: Assessment of fluency and adequacy in translations. Technical report.

Lawrence E. Marks. 1974. *Sensory Processes: The New Psychophysics*. Academic Press, New York.

Saif M. Mohammad and Peter D. Turney. 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text* (caa, 2010), pages 26–34.

Saif M. Mohammad and Peter D. Turney. forthcoming. Crowdsourcing a wordemotion association lexicon. *Computational Intelligence*.

Diego Molla and Maria Elena Santiago-Martinez. 2011. Development of a corpus for evidence based medicine summarisation. In *Proceedings of the Australasian Language Technology Association Workshop 2011*, pages 86–94, Canberra, Australia, December.

Finn rup Nielsen. 2011. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In Matthew Rowe, Milan Stankovic, Aba-Sah Dadzie, and Mariann Hardey, editors, *Proceedings, 1st Workshop on Making Sense of Microposts (#MSM2011): Big things come in small packages*, pages 93–98, Heraklion, Crete, Greece, May.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL05)*, pages 115–124.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii, October. Association for Computational Linguistics.

Philip J. Stone, Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilvie. 1966. *General Inquirer: A Computer Approach to Content Analysis*. The MIT Press.

Carlo Strapparava and Alessandro Valitutti. 2004. WordNet-Affect: an affective extension of WordNet. In *Proceedings of the fourth international conference on Language Resources and Evaluation (LREC 2004)*, pages 1083–1086.

Fangzhong Su and Katja Markert. 2008. From words to senses: A case study of subjectivity recognition. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 825–832, Manchester, UK, August. Coling 2008 Organizing Committee.

Julie Elizabeth Weeds. 2003. *Measures and Applications of Lexical Distributional Similarity*. Ph.D. thesis, University of Sussex.

Simon Whitehead and Lawrence Cavedon. 2010. Generating shifting sentiment for a conversational agent. In *Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text* (caa, 2010), pages 89–97.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2–3):165–210.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, Vancouver, Canada.

Yiming Yang and Jan O. Pedersen. 1997. A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning*, pages 412–420, San Francisco, USA. Morgan Kaufmann Publishers Inc.