FURTHER ANALYSIS OF LATENT AFFECTIVE MAPPING FOR NATURALLY EXPRESSIVE SPEECH SYNTHESIS

Jerome R. Bellegarda
Speech & Language Technologies
Apple Inc., Cupertino, California 95014, USA

ABSTRACT
An essential step in the generation of expressive speech synthesis is the automatic detection and classification of emotions most likely to be present in textual input. At last Interspeech, we introduced latent affective mapping, a new emotion analysis approach which leverages two separate levels of semantic information: one that encapsulates the foundations of the domain considered, and one that specifically accounts for the overall affective fabric of the language [1]–[2]. The ensuing framework exposes the emergent relationship between these two levels in order to advantageously inform the emotion classification process. This paper presents further validation of latent affective mapping, as well as a detailed analysis of its behavior given the richer task of assigning emotional labels to each individual word.

1. INTRODUCTION
To impart emotional quality to synthetic speech, it is necessary to solve two complementary sub-problems: (i) the appropriate emotion must be identified from the given text, and (ii) the corresponding modifications must be effected in speech generation [3]. A number of techniques, often closely tied to the synthesis framework adopted, have been explored to address (ii): cf., e.g., [4]. Comparative analysis of these approaches is beyond the scope of this paper, although it is worth noting that while they may differ in detail, they are largely similar in spirit to the system presented here. The focus here is on the detection part: the reverse task of identifying the primary emotional label(s) for a given text.

We have recently introduced a new framework for emotion detection and classification [1]–[2], based on two separate levels of semantic information: one that encapsulates the foundations of the domain considered, and one that specifically accounts for the overall affective fabric of the language. This approach leverages the latent topicality of two distinct corpora, as uncovered by a global latent semantic mapping (LSM) analysis [7]. The emergent relationship between the two levels is then exploited to expose the desired connection across all terms and emotional categories. Because this connection automatically takes into account the influence of each training corpus, it is more encompassing than that based on the relatively few “affective terms” typically considered in conventional processing. As a result, it effectively bypasses the need for any explicit external information.

In [1]–[2], latent affective mapping was compared to standard approaches based on affective weights and similar expert knowledge. In this paper, we present further validation of the framework, as well as a detailed analysis of its behavior given the resulting richer task of assigning emotional labels to each individual word.

2. LATENT AFFECTIVE MAPPING
Let $T_j$, $|T_j| = N_1$, be a collection of training texts (be they sentences, paragraphs, or documents) reflecting the domain of interest, and $V_1$, $|V_1| = M_1$, the associated set of all words (possibly augmented with some strategic word pairs, triplets, etc., as appropriate) observed in this collection. Similarly, let $T_2$, $|T_2| = N_2$, represent a separate collection of mood-annotated texts (again sentences, paragraphs, or documents), representative of the desired categories of emotions (such as joy and sadness). We denote by $V_2$, $|V_2| = M_2$, the associated set of words or expressions observed in this collection.

Latent affective mapping proceeds as illustrated in Fig. 1. First, we use LSM to encapsulate the semantic information present in the domain corpus $T_1$. This is done by constructing a $(M_1 \times N_1)$ matrix $W_1$, whose elements $w_{ij}$ suitably reflect the extent to which each word $w_i \in V_1$ appeared in each text $t_j \in T_1$. From [1], $w_{ij}$ is given...
Latent affective mapping was evaluated using the news headlines dataset originally put together for the SemEval 2007 task on “Affective Text” [8]. Headlines are normally written by creative people with the intention to “provoke” emotions and consequently attract the readers’ attention. These characteristics make them attractive for use in an automatic emotion recognition setting, as a variety of emotional overtones tend to be present despite the brevity of such sentences. The test data accordingly consisted of 1,250 short news headlines extracted from news web sites (such as Google news, CNN) and/or newspapers, and annotated along $L = 6$ emotions (ANGER, DISGUST, FEAR, JOY, SADNESS, and SURPRISE) by different evaluators. The reader is referred to [10] for detailed information on data annotation, including studies on inter-annotator agreement.

Baseline results were obtained with four different implementations: (i) a simple word accumulation system, which annotates the emotions in a text based on the presence of words from the WordNet-Affect lexicon [8]; and (ii) three systems based on standard latent semantic analysis (LSA), where each emotion is conveyed based on a specific word only (e.g., JOY), or the word plus its WordNet synset, or the word plus all WordNet synsets labelled with that emotion in WordNet-Affect (cf. [8]). In all three cases, the reference corpus used for LSA processing was the Wall Street Journal text collection, comprising about 86,000 articles.

Latent affective mapping was set up as follows. For the “domain” corpus, we selected a collection of about $N_1 = 8,500$ relatively short English sentences (with a vocabulary of roughly $M_1 = 12,000$ words) originally compiled for the purpose of building a concatenative text-to-speech voice. Though not completely

Similarly, LSM can be used a second time, in order to encapsulate the semantic information present in the affective corpus $T_2$ into its own, separate latent semantic space. This makes the affective information more immune to any distribution mismatch between the two corpora, which in turn results in an improved ability to resolve subtle distinctions between emotional connotations. In this case, the mood-annotated texts from the affective corpus $T_2$ are used to construct a $(M_2 \times L)$ matrix $W_2$, whose elements $w_{p, \ell}$ suitably reflect the extent to which each word or expression $w_p \in V_2$ appeared in each affective category $c_\ell$, $1 \leq \ell \leq L$. This leads to definitions analogous to (1), albeit with domain texts replaced by affective categories. We then perform the SVD of $W_2$ in a similar vein as (2). With analogous notation, the resulting decomposition encapsulates a mapping between the set of words $w_p$ and categories $c_\ell$ and (after appropriate scaling by the singular values) the set of $R_2$-dimensional vectors $y_{2,p} = u_{2,p}S_2$ and $z_{2,\ell} = v_{2,\ell}S_2$ in the affective space $\mathcal{L}_2$ spanned by the associated singular vectors.

Thus, each vector $z_{2,\ell}$ can be viewed as the centroid of an emotion in $\mathcal{L}_2$, or, said another way, an affective anchor in the affective space. Since their relative positions reflect a parsimonious encoding of the affective annotations observed in the emotion corpus, these affective anchors now properly take into account any accidental skew in the distribution of words which contribute to them. All that remains is to map them back to the domain space. In latent affective embedding [2], this is done on the basis of entities that are common to both the affective space and the domain space. In the spirit of the morphemic transformation proposed in [9], we derive a cross-space transformation that stipulates how to leverage the observed affective anchors $z_{2,\ell}$ in the affective space to obtain an estimate of the unobserved affective anchors $z_{1,\ell}$ in the domain space, for $1 \leq \ell \leq L$. The reader is referred to [2] for complete details.

$3. \ \text{EXPERIMENTAL RESULTS}$

Latent affective mapping was evaluated using the news headlines dataset originally put together for the SemEval 2007 task on “Affective Text” [8]. Headlines are normally written by creative people with the intention to “provoke” emotions and consequently attract the readers’ attention. These characteristics make them attractive for use in an automatic emotion recognition setting, as a variety of emotional overtones tend to be present despite the brevity of such sentences. The test data accordingly consisted of 1,250 short news headlines extracted from news web sites (such as Google news, CNN) and/or newspapers, and annotated along $L = 6$ emotions (ANGER, DISGUST, FEAR, JOY, SADNESS, and SURPRISE) by different evaluators. The reader is referred to [10] for detailed information on data annotation, including studies on inter-annotator agreement.

Baseline results were obtained with four different implementations: (i) a simple word accumulation system, which annotates the emotions in a text based on the presence of words from the WordNet-Affect lexicon [8]; and (ii) three systems based on standard latent semantic analysis (LSA), where each emotion is conveyed based on a specific word only (e.g., JOY), or the word plus its WordNet synset, or the word plus all WordNet synsets labelled with that emotion in WordNet-Affect (cf. [8]). In all three cases, the reference corpus used for LSA processing was the Wall Street Journal text collection, comprising about 86,000 articles.

Latent affective mapping was set up as follows. For the “domain” corpus, we selected a collection of about $N_1 = 8,500$ relatively short English sentences (with a vocabulary of roughly $M_1 = 12,000$ words) originally compiled for the purpose of building a concatenative text-to-speech voice. Though not completely
Table 1. Results on SemEval-2007 Test Corpus.

<table>
<thead>
<tr>
<th>Approach Considered</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline Word Accumulation</td>
<td>44.7</td>
<td>2.4</td>
<td>4.6</td>
</tr>
<tr>
<td>LSA (Specific Word Only)</td>
<td>11.5</td>
<td>65.8</td>
<td>19.6</td>
</tr>
<tr>
<td>LSA (With WordNet Synset)</td>
<td>12.2</td>
<td>77.5</td>
<td>21.1</td>
</tr>
<tr>
<td>LSA (With All WordNet Synsets)</td>
<td>11.4</td>
<td>89.6</td>
<td>20.3</td>
</tr>
<tr>
<td>Latent Affective Folding</td>
<td>18.8</td>
<td>90.1</td>
<td>31.1</td>
</tr>
<tr>
<td>Latent Affective Embedding</td>
<td>20.9</td>
<td>91.7</td>
<td>34.0</td>
</tr>
</tbody>
</table>

Fig. 3. Probability Distributions Across Emotions for Headline H1.

Fig. 4. Probability Distributions Across Emotions for Headline H2.

A legitimate question to ask is how much these results are influenced by the particular training setup adopted. To ascertain the matter, we retrained the baseline LSA systems using all possible combinations of training sets considered: (i) Wall Street Journal plus text-to-speech corpus; (ii) Wall Street Journal plus blog corpus; (iii) Wall Street Journal plus both text-to-speech and blog corpora; (iv) text-to-speech corpus only; (v) blog corpus only; (vi) text-to-speech plus blog corpus. Results with (i)-(iii) were not statistically different from each other nor from those reported in Table 1, which underscores the dominance of the large corpus relative to the two smaller corpora. Results with (iv)-(vi) turned out to be uniformly worse than those reported in Table 1. While this could be construed as a consequence of a smaller corpus size, it is equally likely to be a reflection of the ensuing mismatch between the feature space and the affective categories, since the latter are still obtained from an external reference lexical database. This may explain why the latent affective approach is seemingly less sensitive to the size of the training sets selected.

4. ILLUSTRATIVE CASE STUDIES

To illustrate of the behavior of latent affective mapping, Figs. 3–6 show the probability distribution obtained across the 6 emotions for the following four headlines:

- H1. Steelers’ Roethlisberger has concussion
- H2. Messi makes Barcelona squad return
- H3. Reunited Police to start world tour in May
- H4. Swingers more common than most think
Headline H1 was chosen for its strong alignment with one emotion (SADNESS), primarily prompted by one word (“concussion”). Fig. 3 shows the associated annotated distribution (black solid line), along with the distributions obtained with the best LSA baseline (with WordNet synset, red dotted line), latent affective folding (purple dashed line) and latent affective embedding (blue dot-dashed line). While in all cases SADNESS prevails, the two latent affective techniques approximate the true distribution better than baseline LSA.

A similar observation can be made regarding Headline H2, selected to show what happens in the absence of a single dominant emotion. Fig. 4 displays the same four distributions as seen in Fig. 3, and once again latent affective mapping proves more effective at following the true distribution. Note, however, that in this more difficult case the exact ranking of the three most salient emotions (JOY, SADNESS, and SURPRISE) is not preserved.

Like H1, Headline H3 displays a strong alignment with one emotion (JOY). This time, however, latent affective mapping is less successful in approximating the true distribution. As shown in Fig. 5, it ranks FEAR before JOY. This is likely due to the polysemic word “police,” which in its alternative sense of civil force indeed tends to be associated with topics like crime and violence.

A similar observation can be made for Headline H4, for which the emotion distribution is broader. Fig. 6 shows a spurious peak at DISGUST, which likely reflects an unintended bias in the training data when it comes to the word “swingers.” This kind of phenomenon is difficult to avoid in emotional corpora, since they normally contain strong individual opinions. Note, however, that both latent affective mapping techniques rank the five other emotions correctly, which is not the case for the LSA baseline. Overall the underlying trade-off thus appears to be favorable to the latent affective approach.

5. CONCLUSION

We have further characterized the behavior of the latent affective mapping framework introduced in [1]–[2] for the data-driven analysis of emotion in text. This framework separately encapsulates both the foundations of the domain considered and the overall affective fabric of the language, and then exploits the emergent relationship between these two semantic levels of description in order to inform the emotion classification process. Evaluations on the “Affective Text” portion of the SemEval-2007 corpus [8] show that the latent affective approach outperforms both affectively weighted word accumulation and standard LSA solutions based on expert knowledge of “affective terms.” In particular, representative case studies point to a better approximation of the true probability distribution across the range of emotions considered. This bodes well for the general deployability of latent affective mapping as an essential pre-processing step in the generation of naturally expressive speech synthesis.

6. REFERENCES