X₀, X₁: Two non-parallel corpora of different “styles”
x⁽i⁾: iᵗʰ sentence of style t
y⁽i⁾: style vector for style t
\( E: X \times Y \to Z \quad G: Y \times Z \to X \)
\( \mathbf{x} = (G(y⁽i⁾, x⁽i⁾), E(x⁽i⁾, y⁽i⁾)). 

**Application** Generating textual paraphrases with modified attributes or stylistic properties (politeness, formality, etc.), benefiting dialogue, writing assistance, etc.; See Pang (2015) for more applications.

**Lack of parallel corpora** ⇒ Need unsup learning criteria and unsup evaluation metrics

**Three goals**
- Correct transfer (by classifier), semantic similarity, fluency
- Datasets Yelp (positive vs. negative), Literature (Dickens vs. Modern)

**Improvements to Eval Metrics**

**3 (Sim) Semantic similarity**

**Def** (i) Embed sentences by avg word embeddings (GloVe, 300d) weighted by idf; (ii) Sim is the avg of the cos sim over all original/ transferred sentence pairs

- Also tried METEOR (large Spearman’s correlation with Sim)
- Simplicity ⇒ efficient & good for widespread adoption

**4 PP (fluency)**

**Def** Measured by language model trained on concat of two corpora

- PP is distinct from fluency, but correlated
- Punished abnormally small PP below

**5 Summarizing Acc, Sim, PP into one single number called GM**

\( GM(y) = (\frac{100}{\text{Acc} \cdot \text{Sim} \cdot \text{PP) min} \{0.01, 0.01, 0.01\}}, \frac{\text{Sim}}{0.01}, \frac{\text{PP}}{0.01}, \frac{\text{Acc}}{0.01}) \)

- Sampled 300 pairs of transferred sentences and asked annotators which one is better
- Training params in GM: t’s are trained by 

\( L_{GM}(t) = \text{max}(0, -GM(y) + GM(y)’ - 1) \)

- t = (63, 71, 97, -37) in our experiments

**6 Result (a): Metric Relationships**

**7 Result (b): System-Level Validation**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Models</th>
<th>Transfer quality</th>
<th>Semantic preservation</th>
<th>Fluency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M0 : M2</td>
<td>9.0 &gt; 6.0 &gt; 85.1</td>
<td>1.5 &gt; 23.4 &gt; 73.1</td>
<td>0.15 &gt; 10.4 &gt; 23.9</td>
</tr>
<tr>
<td>Yelp</td>
<td>M0 : M7</td>
<td>9.6 &gt; 14.7 &gt; 75.8</td>
<td>2.5 &gt; 54.5 &gt; 42.9</td>
<td>0.09 &gt; 4.9 &gt; 39.4</td>
</tr>
<tr>
<td></td>
<td>M0 : M6</td>
<td>11.7 &gt; 11.4 &gt; 74.7</td>
<td>16.0 &gt; 16.7 &gt; 65.4</td>
<td>0.01 &gt; 10.3 &gt; 20.0</td>
</tr>
<tr>
<td></td>
<td>M2 : M7</td>
<td>5.8 &gt; 9.3 &gt; 84.8</td>
<td>8.1 &gt; 25.6 &gt; 66.3</td>
<td>0.04 &gt; 14.0 &gt; 26.7</td>
</tr>
<tr>
<td>Literature</td>
<td>M0 : M2</td>
<td>4.2 &gt; 6.7 &gt; 89.2</td>
<td>16.7 &gt; 20.8 &gt; 62.5</td>
<td>0.01 &gt; 40.8 &gt; 13.3</td>
</tr>
<tr>
<td></td>
<td>M0 : M6</td>
<td>15.8 &gt; 13.3 &gt; 70.8</td>
<td>25.0 &gt; 9.2 &gt; 65.8</td>
<td>0.03 &gt; 14.2 &gt; 20.8</td>
</tr>
</tbody>
</table>

Above table: Human judgments b/w transferred sentences from model A and model B

**Summary** Human judgments in line with automatic measures for semantic preservation and fluency

**Textual transfer evaluation + model code: yzpang.me**