Appendices

A. The KL-divergence between induced distributions

We consider the words found by minimising the difference KL-divergences considered in Section 5. Specifically:

$$w_*^{(1)} = \underset{w_i \in \mathcal{E}}{\operatorname{argmin}}\, D_{KL}[\, p(c_j|\mathcal{W}) \,||\, p(c_j|w_i) \,]$$
$$w_*^{(2)} = \underset{w_i \in \mathcal{E}}{\operatorname{argmin}}\, D_{KL}[\, p(c_j|w_i) \,||\, p(c_j|\mathcal{W}) \,]$$

Minimising $D_{KL}[\, p(c_j|\mathcal{W}) \,||\, p(c_j|w_i) \,]$ identifies the word that induces a probability distribution over context words closest to that induced by $\mathcal{W}$, in which probability mass is assigned to c_j *wherever* it is for $\mathcal{W}$. Intuitively, $w_*^{(1)}$ is the word that most closely reflects *all* aspects of $\mathcal{W}$, and may occur in contexts where no word $w_i \in \mathcal{W}$ does.

Minimising $D_{KL}[\, p(c_j|w_i) \,||\, p(c_j|\mathcal{W}) \,]$ finds the word that induces a distribution over context words that is closest to that induced by $\mathcal{W}$, in which probability mass is assigned as broadly as possible but *only* to those c_j to which probability mass is assigned for $\mathcal{W}$. Intuitively, $w_*^{(2)}$ is the word that reflects as many aspects of $\mathcal{W}$ as possible, as closely as possible, but nothing additional, e.g. by having other meaning that $\mathcal{W}$ does not.

A.1. Weakening the paraphrase assumption

For a given word set $\mathcal{W}$, we consider the relationship between embedding sum $\mathbf{w}_\mathcal{W}$ and embedding $\mathbf{w}_*$ for the word $w_* \in \mathcal{E}$ that minimises the KL-divergence (we illustrate with $\Delta_{KL}^{\mathcal{W},w_*}$). Exploring a weaker assumption than D1, tests whether D1 might exceed requirement, and explores the relationship between $\mathbf{w}_*$ and $\mathbf{w}_\mathcal{W}$ as paraphrase error increases.

Theorem 4 (Weak paraphrasing). *For $w_* \in \mathcal{E}, \mathcal{W} \subseteq \mathcal{E}$, if w_* minimises $\Delta_{KL}^{\mathcal{W},w_*} \doteq D_{KL}[\, p(c_j|\mathcal{W}) \,||\, p(c_j|w_*) \,]$, then:*

$$\mathbf{w}_*^\top \hat{\mathbf{c}} \;=\; \mathbf{w}_\mathcal{W}^\top \hat{\mathbf{c}} - \Delta_{KL}^{\mathcal{W},w_*} + \hat{\sigma}^\mathcal{W} - \tau^\mathcal{W} \tag{17}$$

where $\hat{\mathbf{c}} = \mathbb{E}_{j|\mathcal{W}}[\mathbf{c}_j]$, $\hat{\sigma}^\mathcal{W} = \mathbb{E}_{j|\mathcal{W}}[\boldsymbol{\sigma}_j^\mathcal{W}]$ and $\mathbb{E}_{j|\mathcal{W}}[\cdot]$ denotes expectation under $p(c_j|\mathcal{W})$.

Proof.

$$\begin{aligned}\Delta_{KL}^{\mathcal{W},w_*} &= \textstyle\sum_j p(c_j|\mathcal{W}) \log \frac{p(c_j|\mathcal{W})}{p(c_j|w_*)} \\ &\overset{(5)}{=} \mathbb{E}_{j|\mathcal{W}}[\,\textstyle\sum_i \mathrm{PMI}(w_i, c_j) \\ &\qquad - \mathrm{PMI}(w_*, c_j) + \boldsymbol{\sigma}_j^\mathcal{W} - \tau^\mathcal{W}] \\ &= \mathbb{E}_{j|\mathcal{W}}[\mathbf{w}_\mathcal{W}^\top \mathbf{c}_j - \mathbf{w}_*^\top \mathbf{c}_j] + \hat{\sigma}^\mathcal{W} - \tau^\mathcal{W} \qquad \square\end{aligned}$$

Thus, the weaker paraphrase relationship specifies a hyperplane containing $\mathbf{w}_*$ and so does not uniquely define $\mathbf{w}_*$ (as under D1) and cannot explain the observation of embedding addition for paraphrases (as suggested by Gittens et al. (2017)). A similar result holds for $\Delta_{KL}^{w_*,\mathcal{W}}$. In principle, Thm 4 could help locate embeddings of words that more loosely paraphrase $\mathcal{W}$, i.e. with increased paraphrase error.

B. Proof of Lemma 1

Lemma 1. *For any word $w_* \in \mathcal{E}$ and word set $\mathcal{W} \subseteq \mathcal{E}$, $|\mathcal{W}| < l$:*

$$\mathrm{PMI}_* = \sum_{w_i \in \mathcal{W}} \mathrm{PMI}_i + \boldsymbol{\rho}^{\mathcal{W},w_*} + \boldsymbol{\sigma}^\mathcal{W} - \tau^\mathcal{W}\mathbf{1}\,, \tag{5}$$

where $\mathrm{PMI}_\bullet$ is the column of $\mathbf{PMI}$ corresponding to $w_\bullet \in \mathcal{E}$, $\mathbf{1} \in \mathbb{R}^n$ is a vector of 1s, and error terms $\boldsymbol{\sigma}_j^\mathcal{W} = \log \frac{p(\mathcal{W}|c_j)}{\prod_i p(w_i|c_j)}$ and $\tau^\mathcal{W} = \log \frac{p(\mathcal{W})}{\prod_i p(w_i)}$.

Proof.

$$\begin{aligned}&\mathrm{PMI}(w_*, c_j) - \sum_{w_i \in \mathcal{W}} \mathrm{PMI}(w_i, c_j) \\ &\quad= \log \frac{p(w_*|c_j)}{p(w_*)} - \log \prod_{w_i \in \mathcal{W}} \frac{p(w_i|c_j)}{p(w_i)} \\ &\quad= \log \frac{p(w_*|c_j)}{\prod_\mathcal{W} p(w_i|c_j)} - \log \frac{p(w_*)}{\prod_\mathcal{W} p(w_i)} \\ &\qquad\quad + \log \frac{p(\mathcal{W}|c_j)}{p(\mathcal{W}|c_j)} + \log \frac{p(\mathcal{W})}{p(\mathcal{W})} \\ &\quad= \log \frac{p(w_*|c_j)}{p(\mathcal{W}|c_j)} - \log \frac{p(w_*)}{p(\mathcal{W})} \\ &\qquad\quad + \log \frac{p(\mathcal{W}|c_j)}{\prod_\mathcal{W} p(w_i|c_j)} - \log \frac{p(\mathcal{W})}{\prod_\mathcal{W} p(w_i)} \\ &\quad= \log \frac{p(c_j|w_*)}{p(c_j|\mathcal{W})} + \log \frac{p(\mathcal{W}|c_j)}{\prod_\mathcal{W} p(w_i|c_j)} \\ &\qquad\quad - \log \frac{p(\mathcal{W})}{\prod_\mathcal{W} p(w_i)} \\ &\quad= \boldsymbol{\rho}_j^{\mathcal{W},w_*} \;+\; \boldsymbol{\sigma}_j^\mathcal{W} \;-\; \tau^\mathcal{W}\,,\end{aligned}$$

where, unless stated explicitly, products are with respect to all w_i in the set indicated. $\square$

Introduced terms are highlighted to show their evolution within the proof. At the step where terms are introduced, the existing error terms have no statistical meaning. This is resolved by introducing terms to which both error terms can be meaningfully related, through paraphrasing and independence.

C. Proof of Lemma 2

Lemma 2. *For any word sets $\mathcal{W}$, $\mathcal{W}_* \subseteq \mathcal{E}$, $|\mathcal{W}|$, $|\mathcal{W}_*| < l$:*

$$\sum_{w_i \in \mathcal{W}_*} \mathrm{PMI}_i = \sum_{w_i \in \mathcal{W}} \mathrm{PMI}_i + \boldsymbol{\rho}^{\mathcal{W},\mathcal{W}_*} + \boldsymbol{\sigma}^{\mathcal{W}} - \boldsymbol{\sigma}^{\mathcal{W}_*} - (\tau^{\mathcal{W}} - \tau^{\mathcal{W}_*})\mathbf{1}\,. \tag{10}$$

Proof.

$$\begin{aligned}
&\sum_{w_i \in \mathcal{W}_*} \mathrm{PMI}(w_i,c_j) - \sum_{w_i \in \mathcal{W}} \mathrm{PMI}(w_i,c_j) \\
&= \log \prod_{w_i \in \mathcal{W}_*} \frac{p(w_i|c_j)}{p(w_i)} - \log \prod_{w_i \in \mathcal{W}} \frac{p(w_i|c_j)}{p(w_i)} \\
&= \log \frac{\prod_{\mathcal{W}_*} p(w_i|c_j)}{\prod_{\mathcal{W}} p(w_i|c_j)} - \log \frac{\prod_{\mathcal{W}_*} p(w_i)}{\prod_{\mathcal{W}} p(w_i)} \\
&\quad + \log \frac{p(\mathcal{W}_*|c_j)}{p(\mathcal{W}_*|c_j)} + \log \frac{p(\mathcal{W}_*)}{p(\mathcal{W}_*)} \\
&\quad + \log \frac{p(\mathcal{W}|c_j)}{p(\mathcal{W}|c_j)} + \log \frac{p(\mathcal{W})}{p(\mathcal{W})} \\
&= + \log \frac{p(\mathcal{W}_*|c_j)}{p(\mathcal{W}|c_j)} - \log \frac{p(\mathcal{W}_*)}{p(\mathcal{W})} \\
&\quad + \log \frac{\prod_{\mathcal{W}_*} p(w_i|c_j)}{p(\mathcal{W}_*|c_j)} - \log \frac{\prod_{\mathcal{W}_*} p(w_i)}{p(\mathcal{W}_*)} \\
&\quad + \log \frac{p(\mathcal{W}|c_j)}{\prod_{\mathcal{W}} p(w_i|c_j)} - \log \frac{p(\mathcal{W})}{\prod_{\mathcal{W}} p(w_i)} \\
&= + \log \frac{p(c_j|\mathcal{W}_*)}{p(c_j|\mathcal{W})} \\
&\quad + \log \frac{p(\mathcal{W}|c_j)}{\prod_{\mathcal{W}} p(w_i|c_j)} - \log \frac{p(\mathcal{W}_*|c_j)}{\prod_{\mathcal{W}_*} p(w_i|c_j)} \\
&\quad - \log \frac{p(\mathcal{W})}{\prod_{\mathcal{W}} p(w_i)} + \log \frac{p(\mathcal{W}_*)}{\prod_{\mathcal{W}_*} p(w_i)} \\
&= \boldsymbol{\rho}_j^{\mathcal{W},\mathcal{W}_*} + \boldsymbol{\sigma}_j^{\mathcal{W}} - \boldsymbol{\sigma}_j^{\mathcal{W}_*} - (\tau^{\mathcal{W}} - \tau^{\mathcal{W}_*})\,,
\end{aligned}$$

where, unless stated explicitly, products are with respect to all w_i in the set indicated. □

The proof is analogous to that of Lem 1, with more terms added (as highlighted) to an equivalent effect. A key difference to single-word (or *direct*) paraphrases (D1) is that the paraphrase is between two word sets $\mathcal{W}$ and $\mathcal{W}_*$ that need not correspond to any single word. The paraphrase error $\boldsymbol{\rho}^{\mathcal{W},\mathcal{W}_*}$ compares the induced distributions of the two sets, following the same principles as direct paraphrasing, but with perhaps less interpretatability.

D. Alternate Proof of Corollary 2.1

Corollary 2.1. *For any words $w_x, w_{x^*} \in \mathcal{E}$ and word sets $\mathcal{W}^+, \mathcal{W}^- \subseteq \mathcal{E}$, $|\mathcal{W}^+|, |\mathcal{W}^-| < l-1$:*

$$\mathbf{w}_{x^*} = \mathbf{w}_x + \mathbf{w}_{\mathcal{W}^+} - \mathbf{w}_{\mathcal{W}^-} + \mathbf{C}^\dagger(\boldsymbol{\rho}^{\mathcal{W},\mathcal{W}_*} + \boldsymbol{\sigma}^{\mathcal{W}} - \boldsymbol{\sigma}^{\mathcal{W}_*} - (\tau^{\mathcal{W}} - \tau^{\mathcal{W}_*})\mathbf{1}), \tag{11}$$

where $\mathcal{W} = \{w_x\} \cup \mathcal{W}^+$, $\mathcal{W}_* = \{w_{x^*}\} \cup \mathcal{W}^-$.

Proof.

$$\begin{aligned}
&\mathrm{PMI}(w_{x^*},c_j) - \mathrm{PMI}(w_x, c_j) \\
&= \log \frac{p(c_j|w_{x^*})}{p(c_j|w_x)} + \log \prod_{w_i \in \mathcal{W}^+} \frac{p(c_j|w_i)}{p(c_j|w_i)} + \log \prod_{w_i \in \mathcal{W}^-} \frac{p(c_j|w_i)}{p(c_j|w_i)} \\
&= \sum_{w_i \in \mathcal{W}^+} \log p(c_j|w_i) - \sum_{w_i \in \mathcal{W}^-} \log p(c_j|w_i) + \log \frac{\prod_{\mathcal{W}_*} p(c_j|w_i)}{\prod_{\mathcal{W}} p(c_j|w_i)} \\
&= \sum_{w_i \in \mathcal{W}^+} \mathrm{PMI}(w_i, c_j) - \sum_{w_i \in \mathcal{W}^-} \mathrm{PMI}(w_i, c_j) + \log \frac{\prod_{\mathcal{W}_*} p(w_i|c_j) \prod_{\mathcal{W}} p(w_i)}{\prod_{\mathcal{W}} p(w_i|c_j) \prod_{\mathcal{W}_*} p(w_i)} \\
&= \sum_{w_i \in \mathcal{W}^+} \mathrm{PMI}(w_i, c_j) - \sum_{w_i \in \mathcal{W}^-} \mathrm{PMI}(w_i, c_j) \\
&\quad + \log \frac{p(c_j|w_{x^*}, W^-)}{p(c_j|w_x,\ W^+)} \\
&\quad + \log \frac{\prod_{\mathcal{W}_*} p(w_i|c_j)}{p(w_{x^*}, W^-|c_j)} \frac{p(w_x, W^+|c_j)}{\prod_{\mathcal{W}} p(w_i|c_j)} \\
&\quad - \log \frac{\prod_{\mathcal{W}_*} p(w_i)}{p(w_{x^*}, W^-)} \frac{p(w_x, W^+)}{\prod_{\mathcal{W}} p(w_i)} \\
&= \sum_{w_i \in \mathcal{W}^+} \mathrm{PMI}(w_i, c_j) - \sum_{w_i \in \mathcal{W}^-} \mathrm{PMI}(w_i, c_j) \\
&\quad + \boldsymbol{\rho}_j^{\mathcal{W},\mathcal{W}_*} + \boldsymbol{\sigma}_j^{\mathcal{W}} - \boldsymbol{\sigma}_j^{\mathcal{W}_*} - (\tau^{\mathcal{W}} - \tau^{\mathcal{W}_*})\,,
\end{aligned}$$

where, unless stated explicitly, products are with respect to all w_i in the set indicated; and $\mathcal{W} = \{w_x\} \cup \mathcal{W}^+$, $\mathcal{W}_* = \{w_{x^*}\} \cup \mathcal{W}^-$ to lighten notation. Multiplying by $\mathbf{C}^\dagger$ completes the proof. □