We consider the words found by minimising the difference whether D1 might exceed requirement, and explores the relations $\Delta$ with $\hat{c}$.

Thus, the weaker paraphrase relationship specifies a hyperplane containing $w_*$ and so does not uniquely define $w_*$. (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 $\Delta_{KL}^W$. In principle, Thm 4 could help locate embeddings of words that more loosely paraphrase $W$, i.e. with increased paraphrase error.

### B. Proof of Lemma 1

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

$$PMI_* = \sum_{w_j \in W} PMI_i + \rho_{w^*, w^*} + \sigma^W - \tau^W 1,$$

where $PMI_*$ is the column of $PMI$ corresponding to $w_* \in \mathcal{E}$, $1 \in \mathbb{R}^n$ is a vector of 1s, and error terms $\sigma^W = \log \prod_{i \in W} p(w_i | c_j)$ and $\tau^W = \log \prod_{i \in W} p(w_i)$.

**Proof.**

$$PMI(w_*, c_j) = \sum_{w_j \in W} PMI(w_i, c_j)$$

$$= \log \frac{p(w_*)}{p(w_i)} - \log \prod_{w_j \in W} \frac{p(w_* | c_j)}{p(w_i | c_j)}$$

$$= \log \frac{p(w_*)}{p(W)} \prod_{i \in W} p(w_i | c_j) - \log \prod_{i \in W} p(w_i)$$

$$= \log \frac{p(w_*)}{p(W)} \prod_{i \in W} p(w_i | c_j) - \log \prod_{i \in W} p(w_i)$$

$$= \rho_{w^*, w^*} + \sigma^W - \tau^W$$

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

**Introductions**

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 $W$, $W_r \subseteq \mathcal{E}$, $|W_r| < l$:

\[
\sum_{w_i \in W_r} \text{PMI}(w_i, c_j) - \sum_{w_i \in W} \text{PMI}(w_i, c_j)
= \log \prod_{w_i \in W_r} \frac{p(w_i|c_j)}{p(w_i)} - \log \prod_{w_i \in W} \frac{p(w_i|c_j)}{p(w_i)}
= \log \frac{\Pi_{w_r} p(w_i|c_j)}{\Pi_{w} p(w_i)} - \log \frac{\Pi_{w_r} p(w_i)}{\Pi_{w} p(w_i)} + \log \frac{p(W_r|c_j)}{p(W|c_j)} + \log \frac{p(W_r)}{p(W)}
= \log \frac{p(W_r|c_j)}{p(W|c_j)} - \log \frac{p(W_r)}{p(W)}
= \rho_j^{w_r} + \sigma_j^{w} - \sigma_j^{w_r} - \tau^{w} - \tau^{w_r},
\]

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

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 $W$ and $W_r$ that need not correspond to any single word. The paraphrase error $\rho_j^{w_r}$ compares the induced distributions of the two sets, following the same principles as direct paraphrasing, but with perhaps less interpretability.

D. Alternate Proof of Corollary 2.1

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

\[
w_{x^*} = w_x + w_{w^+} + w_{w^-} + \mathcal{C}^!(\rho_j^{w^+} + \sigma_j^w - \sigma_j^{w^+} - \tau^w - \tau^{w^+})1.
\]

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

\[
\text{PMI}(w_{x^*}, c_j) - \text{PMI}(w_x, c_j)
= \log \frac{p(c_j|w_{x^*})}{p(c_j|w_x)} + \log \frac{\Pi_{w_r} p(c_j|w_i)}{\Pi_{w} p(c_j|w_i)}
+ \log \frac{\Pi_{w_r} p(c_j|w_i)}{\Pi_{w} p(w_i)} + \log \frac{p(W_r|c_j)}{p(W|c_j)} + \log \frac{p(W_r)}{p(W)}
= \sum_{w_i \in W^+} \log p(c_j|w_i) - \sum_{w_i \in W^-} \log p(c_j|w_i)
+ \log \frac{\Pi_{w_r} p(c_j|w_i)}{\Pi_{w} p(w_i)} + \log \frac{\Pi_{w_r} p(c_j|w_i)}{\Pi_{w} p(w_i)}
= \sum_{w_i \in W^+} \text{PMI}(w_i, c_j) - \sum_{w_i \in W^-} \text{PMI}(w_i, c_j)
+ \log \frac{\Pi_{w_r} p(c_j|w_i)}{\Pi_{w} p(w_i)} + \log \frac{\Pi_{w_r} p(c_j|w_i)}{\Pi_{w} p(w_i)}
= \sum_{w_i \in W^+} \text{PMI}(w_i, c_j) - \sum_{w_i \in W^-} \text{PMI}(w_i, c_j)
+ \rho_j^{w^+} + \sigma_j^w - \sigma_j^{w^+} - \tau^w - \tau^{w^+},
\]

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