Appendix A Used notation

We list the notation used throughout the paper

$\mathbb{V}$: vocabulary of words
$\mathcal{V}$: vocabulary of groups
w, v: a word
F_w: relative frequency of a word w
γ_i, γ_j: a group
$\mathbb{V} \times \Gamma$: set of all possible pairs (w, γ_i)
c_{γ_i}: relative frequency of a group γ_i
γ: an assignment (grouping)
$H(\gamma)$: unigram entropy of a grouping γ
$G\left(c_{\gamma'_j}\right)$: partial enropy of a group γ_i
C: number of groups
$[1, \ldots, C]$ - natural numbers from 1 to C
$\mathbb{N}$ - natural numbers

Appendix B Omitted proofs

Definition 1 (Matroid)**.** *Let* Ω *be a finite set* (universe) *and* $\mathcal{I} \subseteq 2^{\Omega}$ *be a set family* (independent sets)*. A pair* $\mathcal{M} = (\Omega, \mathcal{I})$ *is called* a matroid *if*

1. $\emptyset \in \mathcal{I}$
2. *If* $Q \in \mathcal{I}$ *and* $R \subseteq Q$ *then* $R \in \mathcal{I}$
3. *For any* $Q, R \in I$ *with* $|R| < |Q|$ *there exists* $\{x\} \in Q \setminus R$ *such that* $R \cup \{x\} \in \mathcal{I}$.

Let us denote a family of all grouping sets of $\mathbb{V} \times \mathcal{V}$ as $\mathcal{G}$.

Proof of Lemma **??**. We have to show that $(\mathbb{V} \times \mathcal{V}, \mathcal{G})$ satisfies three condition from the Definition 1.

1. An empty grouping is a grouping.
2. Consider an arbitrary $Q \in \mathcal{G}$ and $R \subset Q$. Since Q defines a grouping, for any $(w, \gamma_i) \in Q$ we have $(w\gamma_j) \notin Q$ for all $\gamma_j \neq \gamma_i$. Therefore, for all $(w, \gamma_i) \in R$ we have $(w\gamma_j) \notin R$ given $\gamma_j \neq \gamma_i$ and thus R defines a grouping as well.
3. Consider two arbitrary $R, Q \in \mathcal{G}$ with $|R| < |Q|$. Let us denote $\{w \in \mathbb{V} : (w, \gamma_i) \in Q \text{ for some } \gamma_i\}$ as $\pi(Q)$. We claim that $|Q| = |\pi(Q)|$. Otherwise, there must exist w such that $(w, \gamma_i), (w, \gamma_j) \in Q$ and $\gamma_i \neq \gamma_j$. However, this is forbidden for a set which defines a grouping. Analogously, $|R| = |\pi(R)|$. Since both R, Q are finite, we have $0 < |Q \setminus R| = |\pi(Q)| - |\pi(R)| = |\pi(Q) \setminus \pi(R)|$. Consider an arbitrary $w' \in \pi(Q) \setminus \pi(R)$ and its group $\gamma_{i'}$ in Q; we have $(w', \gamma_{i'}) \in Q \setminus R$. Moreover, since w' is ungrouped by R, we conclude that $R \cup \{(w', \gamma_{i'})\} \in \mathcal{G}$ and finish the proof.

□

Definition 2 (Submodular function)**.** *A function* $f : 2^{\Omega} \to \mathbb{R}$*, where* Ω *is finite, is* submodular *if for any* $X \subseteq Y \subseteq \Omega$ *and any* $x \in \Omega \setminus Y$ *we have*

$$f(X \cup \{x\}) - f(X) \geq f(Y \cup \{x\}) - f(Y).$$

For any non-negative real x and fixed $a > 0$, we denote $-(x+a)\log_2(x+a) + x \log x$ as $L_a(x)$.

Proof of Lemma **??**. First, we show that $H(Q) \geq 0$ for all $Q \subseteq \mathbb{V} \times \mathcal{V}$. By definition, we have $H(\emptyset) = 0$. Consider an arbitrary non-empty $Q \subseteq \mathbb{V} \times \mathcal{V}$. For any $\gamma_i \in \mathcal{V}$ we have

$$0 \leq c_{\gamma_i} = \sum_{\substack{w \in \mathbb{V}: \\ (w, \gamma_i) \in Q}} F_w \leq \sum_{w \in \mathbb{V}} F_w = 1.$$

Therefore, $-c_{\gamma_i} \log c_{\gamma_i} \geq 0$ and

$$\sum_{i=1}^{C} L\left(c_{\gamma_i}\right) \geq 0.$$

Now we establish submodularity. Consider an arbitrary $Q \subseteq \mathbb{V} \times \mathcal{V}$, $R \subset Q$ and any $(w', \gamma_{i'}) \notin Q$. Let $Q' := Q \cup \{(w', \gamma_{i'})\}$, $R' := R \cup \{(w', \gamma_{i'})\}$. We need to show that

$$H(R') - H(R) \geq H(Q') - H(Q). \tag{1}$$

Let us denote the frequency of the unigram γ_j in Q, Q' as $c_{\gamma_j}(Q)$, $c_{\gamma_j}(Q')$. Since Q and Q' differ only in the group $\gamma_{i'}$ we have

$$\begin{aligned} &H(Q') - H(Q) = \\ &- c_{\gamma_{i'}}(Q') \log c_{\gamma_{i'}}(Q) + c_{\gamma_{i'}}(Q) \log c_{\gamma_{i'}}(Q) \end{aligned} \tag{2}$$

Similarly, (2) holds for $H(R') - H(R)$. Thus, to proof (1) it is enough to show

$$\begin{aligned} -c_{\gamma_{i'}}(R') \log c_{\gamma_{i'}}(R') + c_{\gamma_{i'}}(R) \log c_{\gamma_{i'}}(R) \geq \\ -c_{\gamma_{i'}}(Q') \log c_{\gamma_{i'}}(Q') + c_{\gamma_{i'}}(Q) \log c_{\gamma_{i'}}(Q) \end{aligned}$$

We have $c_{\gamma'_i}(Q') = c_{\gamma'_i}(Q) + F_{w'}$; therefore, (2) can be rewritten as $L_{F_{w'}}(c_{\gamma_{i'}}(Q))$. Similarly, $c_{\gamma'_i}(R') = c_{\gamma'_i}(R) + F_{w'}$ hence we need to establish

$$L_{F_{w'}}(c_{\gamma_{i'}}(R)) \geq L_{F_{w'}}(c_{\gamma_{i'}} Q). \tag{3}$$

For any $(w, i') \in R$ we have $(w, i') \in Q$; thus $c_{\gamma_{i'}}(R) < c_{\gamma_{i'}}(Q)$, and (3) follows from the fact that $L_{F_{w'}}(x)$ is monotone decreasing for all non-negative real x. □

Proof of Theorem **??**. By the result (Lee et al., 2009), the Algorithm **??** outputs the map γ' such that

$$\frac{1}{4+4\varepsilon} H(\gamma^*) \leq H(\gamma'). \tag{4}$$

where γ^* is the grouping which achieves largest value of H. We need to show that the approximation guarantee still holds if $\gamma'(w)$ is undefined for some w.

After Step 8, the groupings γ' and γ differ only for the group i_0; thus,

$$H(\gamma) - H(\gamma') = L\left(c_{\gamma_{i_0}}\right) - L\left(c_{\gamma'_{i_0}}\right).$$

Assume that $H(\gamma) - H(\gamma') < 0$. First, there must exist $j \in \mathcal{V}$ such that

$$L\left(c_{\gamma'_{j_0}}\right) \leq \frac{1}{C} H\left(\gamma'\right)$$

and thus for the group i_0 we have

$$L\left(c_{\gamma'_{i_0}}\right) \leq \frac{1}{C} H\left(\gamma'\right) \tag{5}$$

From (5) and $L(x) \geq 0$ we obtain

$$L\left(c_{\gamma_{i_0}}\right) - L\left(c_{\gamma'_{i_0}}\right) \geq -L\left(c_{\gamma'_{i_0}}\right) \geq -\frac{1}{C} H\left(\gamma'\right)$$

hence

$$H(\gamma) \geq \frac{C-1}{C} H\left(\gamma'\right) \geq \frac{C-1}{4C+4\varepsilon C} H(\gamma^*).$$

For a single matroid constrain, the algorithm from (Lee et al., 2009) runs in time $(|\Omega|)^{O(1)}$ where Ω is the universe. In our case, $\Omega = \mathbb{V} \times \mathcal{V}$ hence the running time is $O(C|\mathbb{V}|)^{O(1)}$. The rest of the Algorithm **??** takes $O(C|\mathbb{V}|)^{O(1)}$ steps, thus we obtain the stated running time and finish the proof. □