

Table 4: Statistics of the used datasets, where density denotes the percentages of positive responses.

	#objects	#attributes	density	object token	attribute token
Region-language	165	45	2.59	single or multi-token	single-token
Animal-behavior	354	25	40.3	single-token	multi-token
Disease-symptom	122	33	6.76	single-token	single or multi-token

APPENDIX

SUPPLEMENTAL RESULTS

Computational resources All experiments were conducted on machines equipped with 4 Nvidia A100 GPU. Our method do not require training of MLMs. The extraction of the formal contexts for our datasets is very fast (≤ 60 seconds).

PROOF OF LEMMAS

Lemma 1 (feasibility). *Let $\mathcal{D} = \{w^i\}_{i=1}^N$ be a dataset consisting of data points generated by Abstraction 1, then there exists an identification algorithm $F : \mathcal{D} \rightarrow I$ such that $F(\mathcal{D})$ converges to the ground-truth formal context I_0 as $N \rightarrow \infty$ almost surely, i.e., $F(\mathcal{D}) \rightarrow I \approx I_0$.*

Algorithm 1 A Formal Context Learning Algorithm

Input: A dataset \mathcal{D} , a set of objects G and attributes M
Output: An estimated formal context incidence matrix $I = [0, 1]^{|G| \times |M|}$

initialize $I = \mathbf{0}^{|G| \times |M|}$
for $w \in \mathcal{D}$ **do**
 for $w_i \in w$ **do**
 for $w_j \in w$ **do**
 if $w_i \in G, w_j \in M$ **then**
 $I_{(G_{w_i}, M_{w_j})} = I_{(G_{w_i}, M_{w_j})} + 1$
 end if
 end for
 end for
end for
normalization: $I = \text{normalize}(I)$
return I

Proof. The algorithm 1 is constructed to derive a formal context I from \mathcal{D} . Our target is to demonstrate that I converges to I_0 almost surely when the number of data points $N \rightarrow \infty$. Let Ω denote the sample space, I_N represent the learned formal context from \mathcal{D}_N , and $X_N = d(I_N, I_0)$ denote the random variable indexed by N . We need to establish that $X_N \xrightarrow{\text{a.s.}} 0$.

Let us define $E_N := \{\omega \in \Omega : X_N(\omega) > \epsilon\}$ for $\epsilon > 0$, where ω represents an element of the sample space. Let $Y_{g,m,t} = \begin{cases} 1 & g, m \in x_t \\ 0 & \text{otherwise} \end{cases}$, and

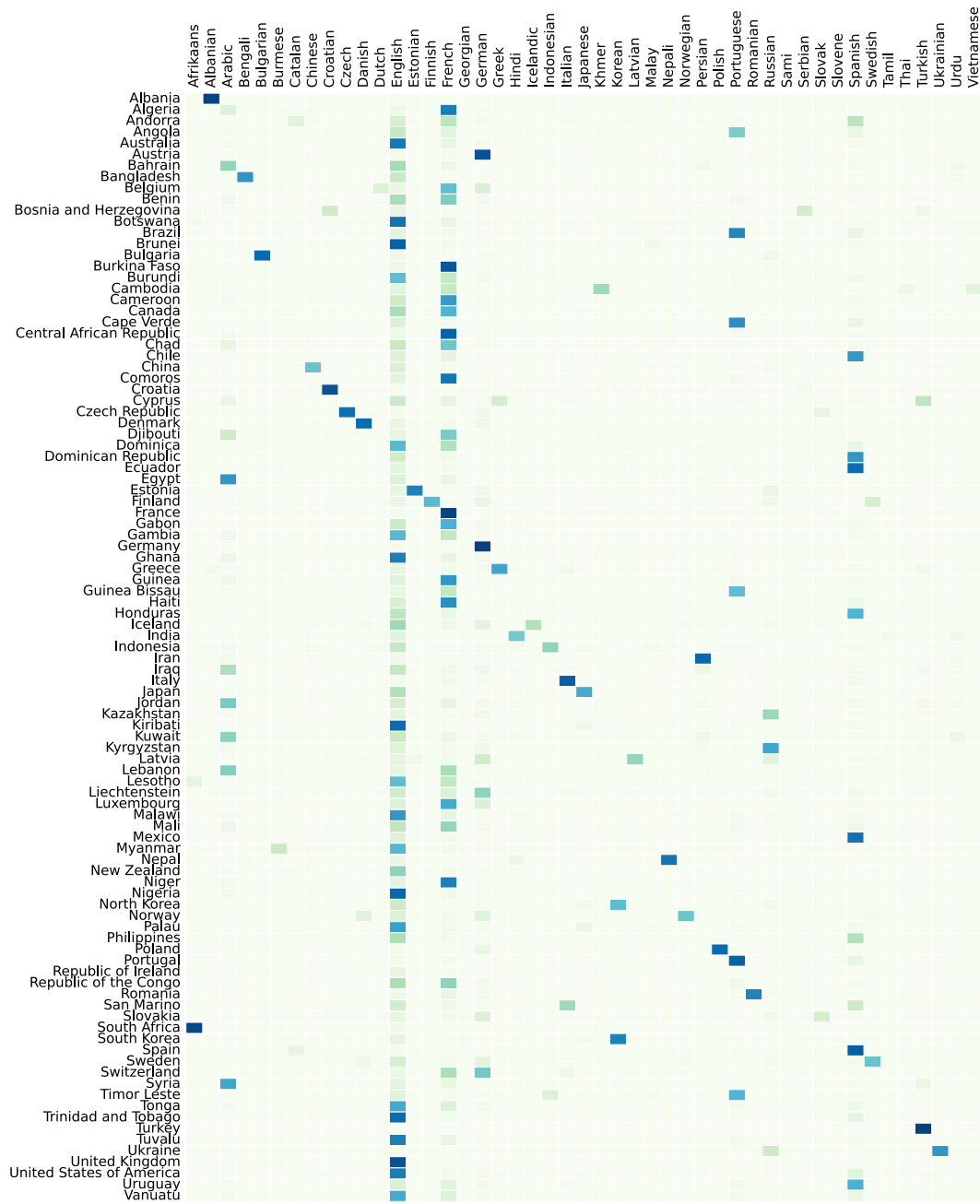


Figure 5: The normalized conditional probability of regions and their official language. The probability is generated by the cloze prompt "[object] is the official language of [attribute]".

$$\begin{aligned}
\text{consider } P(E_N) &= P\left(\sum_{g \in G, m \in M} \left| \frac{1}{N} \sum_{t=1}^N Y_{g,m,t} - I_{0g,m} \right| > \epsilon\right), \quad \text{we have} \\
P(E_N) &\stackrel{(a)}{\leq} P\left(\bigcup_{g \in G, m \in M} \left| \frac{1}{N} \sum_{t=1}^N Y_{g,m,t} - I_{0g,m} \right| > \frac{\epsilon}{q}\right) \\
&\stackrel{(b)}{\leq} \sum_{g \in G, m \in M} P\left(\left| \frac{1}{N} \sum_{t=1}^N Y_{g,m,t} - I_{0g,m} \right| > \frac{\epsilon}{q}\right) \\
&\stackrel{(c)}{\leq} 2q \exp\left(-\frac{2N\epsilon^2}{q^2}\right),
\end{aligned}$$

where inequalities (a) and (b) use the union bound, and (c) applies the Hoeffding's inequality.

By applying the Borel-Cantelli lemma, we have $P(\limsup_{N \rightarrow \infty} E_N) = 0$. Hence, $P(\lim_{N \rightarrow \infty} X_N = 0) = 1$. This means that I converges to I_0 almost surely when $N \rightarrow \infty$. \square

Lemma 2 (role of pattern). *Let $|\text{CMI}_{p_\theta}(g; m|b)|$ denote the conditional MI without latent variables and $\text{CMI}_p(g; m|\mathbf{Z}; b)$ denote the conditional MI with latent variables \mathbf{Z} , we have $|\text{CMI}_{p_\theta}(g; m|b)| - \text{CMI}_p(g; m|\mathbf{Z}; b) \leq 2H(\mathbf{Z}|b)$, where $H(\mathbf{Z}|b)$ is the conditional entropy.*

Proof. Proposition 3 from Zhang & Hashimoto (2021) shows that the dependency between two tokens can be captured by conditional MI. Our lemma can be proved in a similar way by viewing that the objects and attributes are all tokens in the vocabulary. That is $g, m \in V$.

Using the definition of conditional MI, we start with:

$$\text{CMI}_p(g; m|\mathbf{Z}; b) = \text{CMI}_p(g; m|b) - \text{CMI}_p(g; \mathbf{Z}|b) + \text{CMI}_p(g; \mathbf{Z}|m, b)$$

Expanding this, we get:

$$\begin{aligned}
\text{CMI}_p(g; m|\mathbf{Z}; b) &= \text{CMI}_p(g; m|b) + H(\mathbf{Z}|g, b) - H(\mathbf{Z}|b) \\
&\quad + H(\mathbf{Z}|m, b) - H(\mathbf{Z}|g, m, b).
\end{aligned}$$

Now, let's consider the difference:

$$\begin{aligned}
&|\text{CMI}_{p_\theta}(g; m|b)| - \text{CMI}_p(g; m|\mathbf{Z}; b) \\
&= \text{CMI}_p(g; m|b) - \text{CMI}_p(g; m|\mathbf{Z}; b) \\
&= -H(\mathbf{Z}|g, b) + H(\mathbf{Z}|b) - H(\mathbf{Z}|m, b) + H(\mathbf{Z}|g, m, b)
\end{aligned}$$

Next, apply the inequality properties of entropy:

$$\text{CMI}_p(g; m|b) - \text{CMI}_p(g; m|\mathbf{Z}; b) \leq H(\mathbf{Z}|b) + H(\mathbf{Z}|g, m, b)$$

Since entropy is always non-negative, we further have:

$$H(\mathbf{Z}|g, m, b) \leq H(\mathbf{Z}|b)$$

Combining these, we get:

$$\begin{aligned}
|\text{CMI}_{p_\theta}(g; m|b)| - \text{CMI}_p(g; m|\mathbf{Z}; b) &\leq H(\mathbf{Z}|b) + H(\mathbf{Z}|g, m, b) \\
&\leq 2H(\mathbf{Z}|b)
\end{aligned}$$

Therefore, we conclude that:

$$|\text{CMI}_{p_\theta}(g; m|b)| - \text{CMI}_p(g; m|\mathbf{Z}; b) \leq 2H(\mathbf{Z}|b).$$

This completes the proof. \square