1 The ISIC model

The issue-sensitive image captioning model (ISIC model) by Nie, Cohn-Gordon, and Potts, 2020 is defined as follows:

- $\mathcal{I}$ is a set of images; W is the lexicon (a set of words); sequences of words (of arbitrary positive length) are denoted as $\vec{w}$
- the **literal speaker** S_0 is given by a pretrained language model:

  $$S_0(\vec{w} \mid i) \quad \text{[pretrained]}$$

- the **pragmatic listener** L_1 is defined via Bayes rule (where prior probabilities $P(i)$ are assumed to be flat)

  $$L_1(i \mid \vec{w}) \propto P(i)\, S_0(\vec{w} \mid i)$$

- the **issue-sensitive speaker** is defined in terms of three utility components:

  $$\begin{aligned} S_1(\vec{w} \mid i, C) &\propto \exp\left[\alpha((1-\beta)U_1(i, \vec{w}, C) + \beta U_2(i, \vec{w}, C)) - \text{Cost}(i, \vec{w})\right] \text{ , where} \\ U_1(i, \vec{w}, C) &= \log L(C(i) \mid \vec{w}) \\ U_2(i, \vec{w}, C) &= \mathcal{H}(L_1(\cdot \mid \vec{w}, C(i))) \\ \text{Cost}(i, \vec{w}) &= -\log S_0(\vec{w} \mid i) \end{aligned}$$

2 Minimal model setup

To explore the behavior and predictions of this model outside of neural language models, we can look at a minimalized, discrete setup with a ground-truth semantics.

- $\mathcal{I} = F_1 \times \cdots \times F_n$ is a set of objects identified uniquely via a list of n features $F_1, \dots, F_n$. All F_j are non-empty, finite sets which do not contain $\emptyset$ and are mutually disjoint. k_j is the number of feature values in feature set F_j.
- Instead of words and sequences thereof, we look at a set of **messages** $\mathcal{M} = F'_1 \times \cdots \times F'_n$, where $F'_j = F_j \cup \{\emptyset\}$.
- The **meaning function** $\mathcal{B}\colon \mathcal{I} \times \mathcal{M} \to [0; 1]$ is defined as:

  $$\mathcal{B}(i, m) = \begin{cases} 1 & \text{if } \forall j((1 \leq j \leq n) \wedge (m_j \neq \emptyset)) \rightarrow m_j = i_j \\ \epsilon & \text{otherwise.} \end{cases}$$

- An **issue** is a partition on $\mathcal{I}$ derived from a feature. The partition C^j derived from feature j is:

  $$C^i = \left\{\left\{i \in \mathcal{I} \mid i_j = f\right\} \mid f \in F_j\right\}$$

- The **literal speaker** is defined as:

  $$S_0(m \mid i) \propto \mathcal{B}(i, m)$$

- The **parameters** of this setup are n, the list of k_j, α and β and ϵ. The most basic case is $n = 3$ and $k_j = 2$ for all j. ϵ is supposed to be small, e.g., 1×10^{-4}. For extreme (noise-free) predictions, we can set $\epsilon = 0$ and $\alpha \to \infty$.

References

Nie, Allen, Reuben Cohn-Gordon, and Christopher Potts (Nov. 2020). "Pragmatic Issue-Sensitive Image Captioning". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, pp. 1924–1938.