Retreival Augmented Generation

András Kornai

2024 September 25

What we can do

- Build a testbed
- Build a full system
- Do something for Hungarian
- Try to understand the theory better

Readings

- Measuring e2e performance: Krishna et al 2024: FRAMES
- Entity extraction, temporal reasoning: Zhao and Rios 2024
- Build-howto: Church et al 2024
- Eval-howto: https://huggingface.co/learn/cookbook/en/rag_evaluation
- Self-RAG: https://openreview.net/forum?id=hSyW5go0v8
- Classic IR precursors: Xu and Croft 1996

Building a classifier from scratch

Unigram topic model:

$$\binom{l_0 + l_1 + \ldots + l_N}{l_0, l_1, \ldots, l_n} \prod_{i=0}^{N} g_t(w_i)^{l_i} \tag{1}$$

Smoothed with background unigram model:

$$\alpha g_L(w) + (1-\alpha) g_t(w) \tag{2}$$

Prob that topic t emitted doc is:

$$\binom{l_0 + l_1 + \ldots + l_N}{l_0, l_1, \ldots, l_n} \prod_{i=0}^{N} (\alpha g_L(w_i) + (1-\alpha) g_t(w_i))^{l_i} \tag{3}$$

Log probability quotient $\log P(d|t)/P(d|L)$ of doc emitted by t vs L

$$\sum_{i=0}^{N} l_i \log \frac{\alpha g_L(w_i) + (1-\alpha) g_t(w_i)}{g_L(w_i)} \tag{4}$$

Negative evidence: $g_L(w_i) >> g_t(w_i)$

$$\log(\alpha) \sum_{g_L(w_i) >> g_t(w_i)} l_i \tag{5}$$

Positive evidence: $g_L(w_i) << g_t(w_i)$

$$\sum_{g_L(w_i) << g_t(w_i)} l_i \log(\alpha + \frac{(1-\alpha) g_t(w_i)}{g_L(w_i)})$$

Positive evidence simplified

$$\sum_{g_L(w_i) << g_t(w_i)} l_i (\log(1-\alpha) + \log(g_t(w_i)) - \log(g_L(w_i))$$

Notation: *Relevance* $r(w, t)$ of word w to topic t defined as $\log(g_t(w_i)) - \log(g_L(w_i))$. Samples of r for the `alum` topic:

rank	word	$r(w,\texttt{alum})$
1	aluminium	13.4176
2	tonnes	12.9357
3	lme	12.0313
4	alumina	11.9061
1185	though	0.0079206
1186	30	0.00377953
1187	under	0.00100579
1188	second	-0.0146792
1189	7	-0.0207462
1190	with	-0.022297
1316	you	-2.20392
1317	name	-2.96474
1318	country	-2.97375
1319	day	-3.03341

Using relevance to approximate positive evidence

$$\sum_{g_L(w_i) << g_t(w_i)} l_i r(w, t) \tag{6}$$

Log probability quotient $\log P(d|t)/P(d|s)$ of doc emitted by t vs s

$$\log(P(d|t)/P(d|s)) = \log((P(d|t)/P(d|L))/(P(d|s)/P(d|L)))$$

Negative evidence cancels out!

$$\sum_{g_L(w_i) << g_t(w_i)} l_i r(w, t) - \sum_{g_L(w_i) << g_s(w_i)} l_i r(w, s) \tag{7}$$

Bootstrap

- At stage 0, start with some words w_i and assume relevance 1 for these
- Using these, collect positive sample S_0
- At each stage k, $r^{(k)}(w)$ is estimated on positive sample S_{k-1}, and the corpus is reranked to obtain sample S_k
- Tresholds are established by binary search so that the bottom of the sample has about P precision
- Recall is irrelevant (the web is big)
- NER only on positive samples (about 0.25% on Reuters Corpus, 2% on Magyar Hirlap)

English results

- Built many high-precision (0.95+) and medium recall (0.85+) classifiers for a broad range of topics, including petroleum geology and porn (unpublished).
- Trivial infrastructure (requires only aggregation of per-doc wordcounts)
- Low manual effort
- Small S_0 seeds are sufficient
- English seed: *danger, emergency* works well as S_0, manual seeds constructed earlier are roughly equivalent to S_1, little difference by S_2