A Probabilistic Framework for Time Sensitive Search

MPI I at the NTCIR-12 Temporalia-2 Task
Dhruv Gupta  Klaus Berberich
dhgupta@mpi-inf.mpg.de  kberberi@mpi-inf.mpg.de

**Objective**
For a given keyword query, identify a discrete probability distribution over different temporal classes and use this for diversifying documents.

---

**Temporal Intent Disambiguation**
Given classes $C = \text{past}, \text{recent}, \text{future}, \text{atemporal}$ and keyword query $q_{\text{text}}$, estimate $P(C|q_{\text{text}})$.

---

**Step 1. Retrieve pseudo-relevant documents.**

- Time Model - Understand uncertainty behind temporal expressions, e.g., 1930s, by associating lower & upper bounds on begin and end.

### Time Intervals of Interest
- Two step generative model: A time interval is deemed interesting if it's referred frequently by highly relevant documents.

- $P(d_{\text{test}} | q_{\text{test}})$ estimates document likelihood of generating time interval from temporal expressions in document.

- $P([b, e] | d_{\text{time}})$ estimates document likelihood of generating temporal expressions from time interval in document.

- $P([b, e] | q_{\text{test}}) = \sum_{d \in R} P([b, e] | d_{\text{time}}) P(d_{\text{test}} | q_{\text{test}})$

---

**Step 2. Identify interesting time intervals to query ($d_{\text{time}}$) by analyzing the temporal expressions.**

- Generate a discrete probability distribution over the four different temporal classes by analyzing the unit length time interval probability distribution w.r.t. query issue time.

---

**Step 3. Re-rank documents for different temporal categories.**

- $P(q|d) = P(q_{\text{test}}|d_{\text{test}}) \cdot P(q_{\text{time}}|d_{\text{time}})$

- $P(q_{\text{time}}|d_{\text{time}}) = \prod_{[b, e] \in d_{\text{time}}} P([b, e]|d_{\text{time}})$

---

**Step 4. Diversify documents using $q_{\text{time}}$ as aspects.**

- Use the temporal language model to re-rank documents.
- For recent, expand query with time intervals after query issue time.
- For past, expand query with time intervals that lie before query issuing time.
- For future, expand query with time intervals that lie after query issuing time.
- For atemporal, use the pseudo-relevant set of documents.

---

**Temporal Language Model**

- $P([b, e] | q_{\text{test}}) (1-S_{\text{rel}} (1-P(q_{\text{test}}|d_{\text{time}})P([b, e]|d_{\text{time}})))$

---

**Objective Function**

- **Time-Diversity** - The objective is to maximize the probability that the user sees at least one result relevant to her time interval of interest.

- $P([b, e] | q_{\text{test}})$ - Salience of time interval.

- $P(q_{\text{test}}|d_{\text{time}})P([b, e]|d_{\text{time}})$ - Relevance of document for the time interval.

---

**Diversified Documents ($\mathcal{S}$)**

- **D2** - The Economic Historians' View... 1870; 1879; 1880s, early 20th century;...
- **D3** - Don't Blame Wall Street - 2001, 1929; 1920's; early 20th; 90's; 1921; 1923;