بازیابی پرسش در سیستم‌های پرسش و پاسخ
با استفاده از روش‌های مبتنی بر مدل زبانی

دانشجو: شایان پاکزاد
استاد راهنما: دکتر آزاده شاکری
دانشکده مهندسی برق و کامپیوتر، دانشگاه تهران

مقدمه

- **انگیزه:** در حال حاضر با رشد سیستم‌های پرسش و پاسخی همانند Yahoo! Answers و Stack Overflow نیاز به روش‌هایی که دقت بازیابی پرسش را افزایش دهند، احساس می‌شود.
  - بازیابی پرسش به معنای یافتن شبیه‌ترین سوالات موجود در آرشیو با سوالاتی است که در حال حاضر توسط کاربران مطرح می‌شود.
  - این امر دو مزیت دارد:
    ۱. کاربران در سریع‌ترین زمان به هدف خود (پاسخ سوالشان) می‌رسند.
    ۲. آرشیو سیستم ما مملوء از سوالات تکراری نمی‌شود.

- **رویکرد:** ما در این پروژه سعی بر پیشنهاد و پیاده‌سازی یک روش جهت افزایش دقت بازیابی پرسش در این نوع سیستم‌ها داشتیم. بدین منظور روش‌های پیشنهاد شده‌ی مبتنی بر مدل زبانی قبلی را مورد مطالعه قرار دادیم و از ترکیب آن‌ها یک روش جدید را معرفی و پیاده‌سازی کردیم.

- **دستاورد:** در نهایت امر بعد از پیاده‌سازی بهینه روش پیشنهادی و انجام آزمون برروی یک مجموعه داده شاهد افزایش دقت بازیابی پرسش در آن نسبت تعدادی دیگر از روش‌های پیشین بودیم.

روش پیشنهادی

- **روند کلی:** روش‌های بازیابی پرسش مبتنی بر مدل زبانی عموما به شکل زیر اقدام به حل مساله می‌کنند:

ساخت جدول احتمال ترجمه ← **بازیابی پرسش بر اساس مدل آماری**

- **معرفی روش پیشنهادی:**
  - ساخت جدول احتمال ترجمه بوسیله‌ی روش Mutual Information:
    - Mutual Information دو لغت w و u و روابط مرتبط با آن:
      - $I(w;u) = \sum_{X_w=0,1} \sum_{X_u=0,1} p(X_w, X_u) \log \frac{p(X_w,X_u)}{p(X_w)p(X_u)}$
      - $p(X_w = 1) = \frac{c(X_w=1)+2}{N+4}$
      - $p(X_w = 1, X_u = 1) = \frac{c(X_w=1, X_u=1)+1}{N+4}$
    - X_w یک متغیر دودویی: نشان‌دهنده‌ی وجود یا عدم وجود لغت w است.
    - N: تعداد کل سندها (سند = سوال + مجموعه جواب‌هایش)
    - $c(X_w = 1)$: تعداد سندهای شامل لغت w
    - نرمالایز کردن احتمال ترجمه بوسیله رابطه‌ی زیر:
      - $p_{mi}(w|u) = \frac{I(w;u)}{\sum_{w'} I(w';u)}$
    - اضافه کردن پارامتر α برای بهینه کردن احتمال ترجمه به خود:
      - $p_t(w|u) = \begin{cases} \alpha + (1-\alpha)\, p(u|u) & w = u \\ (1-\alpha)\, p(u|u) & w \neq u \end{cases}$
    - بازیابی پرسش بوسیله روابط زیر:
      - $P(q|D) = \prod_{q_i \epsilon Q} Pt(q_i|D)$
      - $P_t(q_i|D) = \frac{\mu}{\mu+|D|} \sum_{t \epsilon D} P_{tran}(q_i|t) P_{ml}(t|D) + \frac{|D|}{\mu+|D|} P_{ml}(q_i|collection)$

پیاده‌سازی و نتایج

- **پیاده سازی:** برای پیاده سازی این روش به دلیل حجم بالای داده‌ی آموزش (مجموعه داده Yahoo! Answers) علاوه بر در نظر گرفتن بهینه‌سازی‌های مربوط به کد تصمیم به در نظر گرفتن طبقه‌هایی خاص از سندها را کردیم. بدین منظور از روابط زیر استفاده کردیم:
  - محاسبه‌ی مدل زبانی هر طبقه بوسیله رابطه‌ی زیر:
    - $p(w|cat_i) = \lambda\, p(w|coll_i) + (1-\lambda)\; p(w|col)$
    - $p(w|d) = \frac{c(w)}{|d|}$
  - طبقه هر سند (پرسش) در داده‌ی آزمون:
    - $cat(q) = argmax_c[\sum_i \log p(q_i|cat_c)]$

- **نتایج:** نمونه هایی از خروجی‌های مربوط به جدول احتمال ترجمه:

جدول ۱: خروجی‌های نمونه جدول احتمال ترجمه برای دو لغت dental و iran (۶ کلمه‌ی اول با بیشترین احتمال)

dental		iran	
dentist	0.05299840	iraq	0.0208048
teeth	0.0446306	iraninian	0.0204473
tooth	0.0302193	muslim	0.0187528
dentists	0.0191190	countries	0.0174564
insurance	0.016908	east	0.0165982
mouth	0.0135634	pakistan	0.01620031

- مقایسه دقت MAP این روش با سایر روش‌ها که برتری این روش را نشان می‌دهد:

جدول ۲: مقایسه دقت روش پیشنهادی با سایر روش‌ها

نام روش	MAP
Okapi	64%
Simple KL Divergence	65%
روش پیشنهادی	67%

جمع بندی

- **خلاصه:** در این پروژه سعی بر پیشنهاد و پیاده‌سازی روشی جدید برای افزایش دقت بازیابی پرسش در سیستم‌های پرسش و پاسخ مبتنی بر انجمن داشتیم که در نهایت امر نیز براساس معیار MAP این امر محقق شد.

- **پیشنهادها:** در رابطه با احتمال ترجمه به خود در صورتی‌که از روشی استفاده شود که در عین بهینه کردن احتمال تبدیل ترجمه به خود کمترین تاثیر را برروی احتمال ترجمه به سایر کلمات بگذارد، می‌تواند دقت را افزایش دهد. همچنین استفاده از ریشه‌یاب‌های مختلف برای ساخت جدول احتمال ترجمه نیز می‌تواند در بهبود دقت موثر باشد.

- **کاربردها:** نتایج این پروژه می‌تواند هم در سیستم‌های پرسش و پاسخ و فروم ها باعث افزایش کیفیت بازیابی پرسش شود و هم می‌تواند از ایده های استفاده شده در آن در سایر مسائل مربوط به حوزه بازیابی اطلاعات مثل Question Answering استفاده شود.

مراجع اصلی

1. **J. Jeon, W. B. Croft, and J. H. Lee, "Finding similar questions in large question and answer archives," in Proceedings of the 14th ACM international conference on Information and knowledge management - CIKM '05, 2005.**
2. **X. Xue, J. Jeon, and W. B. Croft, "Retrieval models for question and answer archives," in Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '08, 2008.**
3. **M. Karimzadehgan and C. Zhai, "Estimation of statistical translation models based on mutual information for ad hoc information retrieval," in Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '10, 2010.**