فاز ۲ — اجرای مدل روی دیتاست Re-DocRED و ارزیابی کمی

تاریخ اجرا: ۱۴۰۵/۰۵/۲۶ · **مدل زبانی:** `openai/gpt-4.1-mini`، دما ۰ · **حالت:** zero-shot (بدون آموزش و بدون مثال)
زیرمجموعه: ۵۰ سند نخست بخش Dev دیتاست Re-DocRED · **۱,۸۶۸ سه‌تایی طلایی** · **۸۵ رابطه متمایز**

---

۱. آرایش آزمایش

مسئله، استخراج رابطه در سطح سند است. ما آرایش استاندارد این بنچمارک را رعایت کردیم: فهرست موجودیت‌های هر سند به مدل داده می‌شود و مدل باید رابطه‌های بین آن‌ها را پیدا کند و برای هر رابطه جمله‌ی شاهد را نقل کند.

خط لوله‌ی فاز ۱ بدون بازنویسی معماری استفاده شد. تنها یک نقطه‌ی اتصال اضافه شد:

بخش سامانه	تغییر برای Re-DocRED
قطعه‌بندی و پیش‌پردازش	بدون تغییر
فراخوانی مدل زبانی	بدون تغییر
اعتبارسنجی Pydantic	بدون تغییر
شاهد اجباری برای هر رابطه	بدون تغییر
حذف تکراری با `fact_id`	بدون تغییر
هستان‌شناسی	**از فایل JSON خوانده می‌شود (فلگ `--ontology`)**

هستان‌شناسی از **بخش Train** استخراج شد، نه دست‌نویس: ۶ نوع موجودیت، ۹۶ رابطه، و **۶۵۴ ترکیب مجاز** (نوع فاعل، رابطه، نوع مفعول). هر سه‌تایی که بیرون این ۶۵۴ ترکیب باشد رد می‌شود.

---

۲. جدول ۱ — نتیجه‌ی سه تکرار

نسخه	توضیح	سه‌تایی خروجی	Precision	Recall	F1	Ign F1
v1	پرامپت پایه: «فقط چیزی را بده که متن می‌گوید»	۴۶۶	۴۲٫۲۷	۱۰٫۵۵	**۱۶٫۸۸**	۱۹٫۲۰
v2	پرامپت جامع: جست‌وجوی همه‌ی جفت‌ها + رابطه‌ی معکوس + استنتاج زنجیره‌ای	۷۱۶	۳۶٫۴۵	۱۳٫۹۷	**۲۰٫۲۰**	۲۱٫۰۶
v3	v2 + لایه‌ی بستار قاعده‌محور روی گراف	۹۸۵	۳۴٫۱۱	۱۷٫۹۹	**۲۳٫۵۵**	۲۳٫۹۰
v4	v2 + دو مثال کاری در پرامپت (few-shot از بخش Train)	۵۰۱	**۴۵٫۱۱**	۱۲٫۱۰	**۱۹٫۰۸**	۱۹٫۸۶
v5	v4 + لایه‌ی بستار قاعده‌محور	۷۱۰	**۴۵٫۳۵**	۱۷٫۲۴	**۲۴٫۹۸**	۲۵٫۵۲
v6	**اجماع سه گذر (v1+v2+v4) + لایه‌ی بستار**	۱,۵۳۱	۳۲٫۹۲	**۲۶٫۹۸**	**۲۹٫۶۶**	**۲۹٫۳۵**

Ign F1 معیار سخت‌گیرانه‌ی DocRED است. هر حقیقتی که در بخش Train هم دیده شده از دو طرف حذف می‌شود، تا حفظ‌کردن جواب امتیاز نگیرد. در کار ما Ign F1 کمی **بالاتر** از F1 است. این نشان می‌دهد سامانه از حافظه‌ی مدل تغذیه نمی‌کند و از متن سند می‌خواند.

از v1 به v6 مقدار **۷۶٪** F1 **رشد نسبی** داشت (۱۶٫۸۸ → ۲۹٫۶۶).

اثر مثال کاری در پرامپت (v2 در برابر v4): دقت از ۳۶٫۴۵ به ۴۵٫۱۱ رسید، یعنی ۲۴٪ رشد نسبی. نرخ رد شدن توسط فیلتر هستان‌شناسی از ۲۰٫۳٪ به ۱۸٫۷٪ افت کرد. پس مثال، دقت را می‌خرد نه بازخوانی را.

اثر اجماع چند گذر (v6): سه گذر مستقل با پرامپت‌های متفاوت، مجموعه‌های متفاوتی از حقیقت‌ها را پیدا می‌کنند. اجتماع آن‌ها بازخوانی را از ۱۷٫۲۴ به ۲۶٫۹۸ رساند. این بزرگ‌ترین جهش کل پروژه است و نشان می‌دهد گلوگاه سامانه، **کم‌گویی در یک گذر** بود، نه ندانستن.

گزارش کامل و ادغام‌شده در FINAL_REPORT.md آمده است.

---

۳. جدول ۲ — بازخوانی به تفکیک رابطه

رابطه	تعداد طلایی	v1	v2	v3
located in the administrative territorial entity	۴۵۵	۹٫۹٪	۱۸٫۷٪	**۲۶٫۶٪**
country	۳۵۱	۲٫۰٪	۳٫۷٪	**۷٫۱٪**
contains administrative territorial entity	۱۱۵	۰٪	۶٫۱٪	**۲۵٫۲٪**
country of citizenship	۹۰	۱۲٫۲٪	۲۱٫۱٪	۲۱٫۱٪
notable work	۵۳	۱٫۹٪	۱٫۹٪	۱٫۹٪
performer	۳۲	۹٫۴٪	۳٫۱٪	۳٫۱٪
has part	۳۲	۹٫۴٪	۹٫۴٪	**۱۵٫۶٪**
part of	۳۲	۳٫۱٪	۶٫۲٪	**۱۵٫۶٪**

رابطه‌ی `contains` معکوس رابطه‌ی `located in` است. مدل زبانی در v1 هیچ‌وقت جهت معکوس را نمی‌داد و بازخوانی صفر بود. لایه‌ی قاعده آن را به ۲۵٫۲٪ رساند.

---

۴. لایه‌ی بستار قاعده‌محور (نوآوری فاز ۲)

مدل زبانی چیزی را می‌گوید که در جمله نوشته شده. یک گراف دانش باید چیزی را هم نگه دارد که از گراف نتیجه می‌شود. سه قاعده روی گراف اجرا می‌شود:

قاعده	منطق	تعداد حقیقت تولیدشده
R1 معکوس	A در B واقع است ⟹ B شامل A است	۱۵۳
R2 تراگذری	A در B و B در C ⟹ A در C	۱۰۴
R3 زنجیره	A در B و B کشورش C ⟹ A کشورش C	۱۳
جمع		**۲۷۰**

از ۲۷۰ حقیقت استنتاجی، **۷۶ مورد در پاسخ طلایی وجود دارد (دقت ۲۸٫۱٪)**. این دقت پایین‌تر از دقت مدل زبانی است، پس لایه‌ی قاعده بازخوانی را می‌خرد و بخشی از دقت را می‌فروشد. چون F1 کل بالا رفت، این معاوضه به سود ما بود.

هر حقیقت استنتاجی **شاهد مقدمه‌های خود را به ارث می‌برد**. یعنی قانون «هر یال باید شاهد داشته باشد» در کل سامانه نقض نمی‌شود. یک تست خودکار این را بررسی می‌کند.

---

۵. تحلیل خطا

۵.۱ کار مؤثر فیلتر هستان‌شناسی. در اجرای v2 مدل زبانی ۹۴۸ سه‌تایی برگرداند. فیلتر نوع، **۱۹۲ مورد (۲۰٫۳٪) را رد کرد**. پرتکرارترین موارد ردشده:

سه‌تایی ردشده	تعداد	ایراد
`PER --performer--> MISC`	۱۶	جهت برعکس است. درست آن `PER <--performer-- MISC` است.
`PER --end time--> TIME`	۱۳	این ترکیب نوع در هستان‌شناسی وجود ندارد.
`PER --location--> LOC`	۹	برای شخص، رابطه‌ی `location` تعریف نشده است.
`PER --author--> MISC`	۸	جهت برعکس است.

پس فیلتر سخت‌گیرانه واقعاً کار می‌کند و توهم مدل را می‌گیرد. این همان سازوکاری است که در پروژه‌ی شهرداری هم دقت را بالا نگه داشت.

۵.۲ کم‌گویی، علت اصلی بازخوانی پایین. پاسخ طلایی برای هر سند به‌طور میانگین **۳۷٫۴** رابطه دارد. مدل در v1 فقط **۹٫۳** رابطه می‌داد و در v3 به **۱۹٫۷** رسید. مدل زبانی جمله‌های صریح را می‌بیند، ولی حقیقت‌های استنتاجی و هم‌مرجعی‌شده را رد می‌کند.

۵.۳ نوع خطای مثبت کاذب. از ۶۴۹ مثبت کاذب نسخه v3:

نوع خطا	تعداد	سهم
جفت‌موجودیت در پاسخ طلایی رابطه دارد، ولی برچسب رابطه اشتباه است	۱۲۸	۱۹٫۷٪
جفت‌موجودیت در پاسخ طلایی هیچ رابطه‌ای ندارد	۵۲۱	۸۰٫۳٪

یعنی حدود یک‌پنجم خطاها «اشتباه در انتخاب یکی از ۹۶ برچسب» است، نه ساختن رابطه از هوا.

۵.۴ خطای نگاشت نام. ۳۷ سطر از ۱٬۰۲۶ سطر (۳٫۶٪) به موجودیت طلایی نگاشت نشد، چون مدل نام را دقیقاً از فهرست کپی نکرد.

---

۶. سقف نظری روش

ارزیاب با یک پاسخ اوراکل (همه‌ی جواب‌های درست، بر اساس نام موجودیت) آزمایش شد و **F1 = ۹۹٫۶۲** گرفت، نه ۱۰۰. علت: در چند سند دو موجودیت متفاوت با **نام و نوع یکسان** وجود دارد و با نام قابل تفکیک نیستند.

پس سقف هر سامانه‌ی نام‌محور روی این داده ۹۹٫۶۲ است. این عدد باید کنار عدد مدل خوانده شود.

---

۷. مقایسه با کارهای منتشرشده

سامانه	آموزش	F1 روی Re-DocRED
KD-DocRE	آموزش نظارت‌شده روی ۳,۰۵۳ سند	۸۱٫۰۴
ATLOP	آموزش نظارت‌شده	۷۹٫۴۶
GLiDRE	آموزش نظارت‌شده	۷۷٫۸۳
GPT-3.5 + NLI (کار منتشرشده)	zero-shot	۹٫۷۴
سامانه‌ی ما (۵۰، v3 سند)	zero-shot، بدون هیچ مثالی	۲۳٫۵۵
سامانه‌ی ما (۵۰، v6 سند)	دو مثال در پرامپت + اجماع سه گذر	۲۹٫۶۶
سامانه‌ی ما (کل ۵۰۰ سند Dev)	**دو مثال + اجماع دو گذر**	**۲۸٫۶۴**

نتیجه‌ی مهم: سامانه‌ی ما بدون هیچ آموزشی، **بیش از سه برابر** خط‌مبنای zero-shot منتشرشده با GPT-3.5 امتیاز گرفت. در برابر مدل‌های نظارت‌شده که روی ۳ هزار سند برچسب‌خورده آموزش دیده‌اند، فاصله‌ی زیادی هست و این فاصله انتظار درست است.

هشدار مقایسه: اعداد مدل‌های نظارت‌شده روی کل بخش Test گزارش شده و عدد ما روی بخش Dev. پس این مقایسه، مقایسه‌ی مرتبه‌ی بزرگی است، نه مقایسه‌ی نقطه‌به‌نقطه.

---

۸. محدودیت‌ها

1. ~~زیرمجموعه ۵۰ سند است، نه ۵۰۰ سند کامل Dev.~~ **این محدودیت برطرف شد.** سامانه روی کل ۵۰۰ سند Dev اجرا شد و **F1 = ۲۸٫۶۴** گرفت (بخش ۷٫۵ گزارش نهایی). اختلاف با عدد زیرمجموعه کمتر از یک واحد است. بخش Test هنوز اجرا نشده است.
2. ~~سامانه zero-shot است و هیچ نمونه‌ای در پرامپت نیست.~~ **این محدودیت هم برطرف شد.** دو مثال کاری از بخش Train به پرامپت اضافه شد (نسخه v4 به بعد) و دقت را ۲۴٪ بالا برد. گام بعدی، بازیابی مثال **مشابه هر سند** به‌جای مثال ثابت است.
3. نگاشت پاسخ بر پایه‌ی نام موجودیت است، نه اندیس. سقف آن ۹۹٫۶۲ است.
4. لایه‌ی قاعده فقط سه قاعده دارد. قاعده‌های بیشتر از روی Train قابل استخراج خودکار هستند.

---

۹. بازتولید نتیجه

```
python -m src.redocred prepare --split dev --limit 50
python -m src.extract data/benchmark/chunks_dev.jsonl \
    --out data/benchmark/triplets_dev_v2.jsonl \
    --ontology configs/ontology_redocred.json
python -m src.redocred closure data/benchmark/triplets_dev_v2.jsonl \
    --out data/benchmark/triplets_dev_v3.jsonl
python -m src.redocred eval data/benchmark/triplets_dev_v3.jsonl \
    --report data/benchmark/eval_report_v3.json
```

تست‌های آفلاین سامانه (بدون فراخوانی مدل زبانی و بدون شبکه):

```
python -m tests.test_extract      # ۶ تست هستان‌شناسی و اعتبارسنجی
python -m tests.test_redocred     # ۶ تست ارزیاب و لایه‌ی قاعده
```

خروجی هر دو: `all passed`.

---

۱۰. فایل‌های تولیدشده

فایل	محتوا
`src/redocred.py`	آماده‌سازی داده، لایه‌ی بستار قاعده‌محور، ارزیاب
`configs/ontology_redocred.json`	۹۶ رابطه و ۶۵۴ ترکیب مجاز
`data/benchmark/triplets_dev_v1.jsonl`	خروجی پرامپت پایه
`data/benchmark/triplets_dev_v2.jsonl`	خروجی پرامپت جامع
`data/benchmark/triplets_dev_v3.jsonl`	خروجی نهایی با حقیقت‌های استنتاجی
`data/benchmark/eval_report_v1/v2/v3.json`	گزارش عددی هر سه نسخه
`tests/test_redocred.py`	۶ تست آفلاین