

Sourena jabbari
An Ensemble Event Extraction Method on News
(Springer, 2025)

1 (مسئله اصلی چیست و چرا مهم است؟

در دنیای امروز، حجم زیادی از اطلاعات و تصمیم‌سازی‌ها بر اساس اخبار آنلاین انجام می‌شود. اما اخبار به صورت متن آزاد منتشر می‌شوند و برای تحلیل خودکار (مثلاً تحلیل سیاسی، تحلیل امنیتی، تحلیل بازار و یا پایش بحران‌ها) لازم است اطلاعات به شکل ساختارمند استخراج شود.

یکی از مهم‌ترین وظایف در پردازش زبان طبیعی (NLP) روی اخبار، **استخراج رویداد** (Event Extraction) است. در این مسئله هدف این است که از متن خبر، رخداد‌های مهم مانند موارد زیر استخراج شوند:

- حمله (Attack)
- بازداشت (Arrest)
- توافق (Agreement)
- حادثه طبیعی (Disaster)
- مرگ و تلفات (Death)

و علاوه بر شناسایی نوع رویداد، عناصر مرتبط با آن (Arguments) نیز استخراج شوند، مانند:

- عامل رویداد (Actor)
- هدف رویداد (Target)
- مکان (Location)
- زمان (Time)

مشکل اصلی اینجا است که اخبار معمولاً دارای ویژگی‌های زیر هستند:

- ابهام زبانی (Ambiguity)
- روایت‌های مختلف از یک اتفاق
- حذف بخشی از اطلاعات (مثلاً زمان یا عامل رویداد)
- پیچیدگی در جمله‌بندی
- وجود اطلاعات متناقض یا غیرمستقیم

در نتیجه، مدل‌های استخراج رویداد ممکن است رویدادها را اشتباه استخراج کنند یا با اطمینان پایین پیش‌بینی کنند.

بنابراین، استخراج رویداد از اخبار نه تنها یک مسئله مهم در NLP است، بلکه برای کاربردهای عملی مثل مانیتورینگ اخبار و تحلیل تصمیم‌گیری بسیار حیاتی است.

2) ورودی‌ها و خروجی‌های مدل/سیستم چیست؟

ورودی سیستم

ورودی سیستم پیشنهادی مقاله یک متن خبری است. این متن می‌تواند یک جمله یا یک خبر کامل باشد.

مثال ورودی:

> "Police arrested a suspect in Berlin on Monday after an attack near the station."

خروجی سیستم

خروجی سیستم یک مجموعه از رویدادهای استخراج‌شده است که شامل موارد زیر هستند:

- نوع رویداد (Event Type)

- کلمه یا عبارت محرک رویداد (Trigger)

- آرگومان‌ها (Arguments)

- امتیاز اطمینان (Confidence Score)

نمونه خروجی ساختاریافته:

```
```json
{
 "event_type": "Arrest",
 "trigger": "arrested",
 "arguments": {
 "actor": "Police",
 "location": "Berlin",
 "time": "Monday"
 },
 "confidence_score": 0.82
}
```

---

### ۳) داده مورد استفاده (نوع، منبع، اندازه)

در مقاله بیان شده است که آزمایش‌ها روی مجموعه‌ای از اخبار آنلاین انجام شده است. اما مقاله دیتاست عمومی رسمی ارائه نمی‌دهد و در صفحه مقاله ذکر شده است که:

No datasets were generated or analysed during the current study.

بنابراین، برای پیاده‌سازی و آزمایش عملی، لازم است از یک دیتاست عمومی جایگزین یا داده‌های نمونه استفاده شود.

در پیاده‌سازی prototype این پروژه، برای نمایش ایده مقاله، از یک دیتاست نمونه کوچک شامل 10 متن خبری استفاده شد که رویدادهای اصلی آن به صورت دستی برچسب‌گذاری شدند.

این نمونه‌داده در مسیر زیر قرار دارد:

- demo/sample\_news.csv
- demo/sample\_labels.csv

با اینکه این داده کوچک است، اما برای اثبات عملکرد الگوریتم اعتمادسنجی (Confidence Score) در مقیاس prototype کافی است.

---

## ۴) روش پیشنهادی مقاله به زبان ساده

مقاله یک روش جدید برای استخراج رویداد پیشنهاد می‌دهد که مبتنی بر ترکیب چند مدل (Ensemble) است.

ایده اصلی مقاله این است که اگر چند مدل مختلف روی یک رویداد توافق داشته باشند، احتمال درست بودن آن بیشتر است. همچنین، هر مدل ممکن است خروجی‌هایی تولید کند که قابل اعتماد نباشند، بنابراین مقاله یک معیار خاص معرفی می‌کند تا میزان اعتماد به خروجی‌ها را کمی‌سازی کند.

### ۴-۱) چرا Ensemble ؟

در بسیاری از مسائل NLP، یک مدل واحد ممکن است در برخی شرایط خوب و در برخی شرایط ضعیف عمل کند. اما اگر چند مدل مستقل استفاده شوند و خروجی آن‌ها ترکیب شود، سیستم نهایی معمولاً پایدارتر و دقیق‌تر خواهد بود.

روش‌های Ensemble معمولاً شامل majority vote یا averaging هستند، اما این مقاله نشان می‌دهد که برای استخراج رویداد، باید یک معیار اعتماد دقیق‌تر وجود داشته باشد.

---

## ۵) مفهوم کلیدی مقاله (Confidence Score (CS :

مقاله معیار جدیدی معرفی می‌کند به نام **Confidence Score (CS)**.

Confidence Score یک عدد است که نشان می‌دهد:

- رویداد استخراج‌شده چقدر معتبر است
- مدل چقدر به پیش‌بینی خود اطمینان دارد

برخلاف برخی سیستم‌ها که فقط خروجی رویداد را تولید می‌کنند، این مقاله یک مکانیزم رسمی برای تعیین میزان اعتماد ارائه می‌دهد.

این ویژگی در سیستم‌های عملی بسیار مهم است، زیرا:

- برخی رویدادها ممکن است مبهم باشند
- برخی خبرها ممکن است جعلی یا ناقص باشند
- سیستم عملی باید بداند کدام خروجی قابل استفاده است

---

## ۶) روش محاسبه CS در مقاله

مقاله پیشنهاد می‌دهد که CS بر اساس دو عامل اصلی محاسبه شود:

### 6-1) Prior Confidence

هر مدل به صورت اولیه یک امتیاز اعتماد دارد. این امتیاز می‌تواند از:

- احتمال خروجی مدل
- کیفیت مدل در آموزش
- یا تنظیمات تجربی

به دست آید.

### 6-2) Consistency بین مدل‌ها

اگر مدل‌های مختلف خروجی مشابه تولید کنند، این نشان می‌دهد که رویداد احتمالاً درست است.

مقاله از مفهوم **pairwise consistency** استفاده می‌کند:

- خروجی مدل A با مدل B مقایسه می‌شود
- خروجی مدل A با مدل C مقایسه می‌شود
- ...

اگر توافق بالا باشد، CS افزایش پیدا می‌کند.

---

## ۷) الگوریتم Iterative Update در مقاله

یکی از نوآوری‌های اصلی مقاله این است که CS تنها یک بار محاسبه نمی‌شود، بلکه در یک فرآیند تکراری (Iterative) بهبود داده می‌شود.

در هر مرحله:

- CS قبلی حفظ می‌شود
- سازگاری مدل‌ها محاسبه می‌شود
- CS با یک وزن‌دهی جدید آپدیت می‌شود

این باعث می‌شود سیستم بتواند به صورت پایدارتر، خروجی‌های مطمئن‌تر را تشخیص دهد.

---

## ۸) شبکه‌کد روش مقاله (تقریب ساده)

در این پروژه، الگوریتم مقاله به صورت ساده به شکل زیر پیاده‌سازی شد:

Input: News text

Output: Extracted events + confidence scores

1. Run multiple models:

Model A -> Events\_A

Model B -> Events\_B

Model C -> Events\_C

2. Initialize CS for each model output:

CS = prior\_confidence

3. Repeat for K iterations:  
For each model output event:  
compute similarity with other model outputs  
consistency = average similarity  
 $CS_{new} = \alpha * CS_{old} + (1 - \alpha) * consistency$

4. Aggregate final events:  
merge similar events  
final\_score = average CS of supporting models

در این شبه‌کد:

- similarity: میزان شباهت رویدادها بر اساس نوع رویداد، trigger و arguments است.
- alpha: وزن prior confidence را مشخص می‌کند.

---

## ۹) پیاده‌سازی Prototype در این پروژه

از آنجا که مقاله کد و دیتاست عمومی ارائه نکرده است، مطابق دستور استاد، یک prototype کوچک برای نمایش ایده مقاله ساخته شد.

### ۹-۱) مدل‌های استفاده‌شده در prototype

در پیاده‌سازی این پروژه، سه استخراج‌کننده رویداد (Event Extractor) به عنوان ensemble استفاده شدند:

#### 1. Rule-based Extractor

- تشخیص triggerهای کلیدی مثل attacked, arrested, signed
- استخراج آرگومان‌ها با regex ساده

#### 2. Heuristic Pattern Extractor

- تشخیص الگوهای متفاوت از مدل اول
- استخراج زمان و مکان با patternهای جداگانه

#### 3. Transformer-based NER Extractor (اختیاری)

- استفاده از HuggingFace مدل آماده NER
- استخراج entity ها و تبدیل آن‌ها به arguments

این ساختار باعث می‌شود حتی در صورت نبود اینترنت یا مدل HF، pipeline همچنان قابل اجرا باشد.

## 9-2 محاسبه Confidence Score

برای هر خروجی مدل:

- prior confidence تعیین شد (مثلاً 0.55 برای rules و 0.60 برای heuristic و 0.72 برای HF)
- سپس pairwise consistency محاسبه شد
- الگوریتم iterative update برای 3 iteration اجرا شد

---

## ۱۰) نتایج اصلی و ارزیابی

برای ارزیابی، معیارهای Precision / Recall / F1 در سطح event\_type محاسبه شد.

خروجی‌ها در فایل زیر ذخیره شدند:

- experiments/results.csv

در حالت کلی، انتظار می‌رود:

- مدل‌های تکی ممکن است برخی رویدادها را اشتباه تشخیص دهند
- روش ensemble همراه با CS به دلیل حذف خروجی‌های کم‌اطمینان و تقویت خروجی‌های توافق‌شده، دقت بهتری داشته باشد

این نتیجه با ادعای مقاله نیز همخوانی دارد که ensemble پیشنهادی را بهتر از مدل تکی و ensemble ساده معرفی می‌کند.

---

## ۱۱) محدودیت‌های مقاله و محدودیت‌های prototype

محدودیت‌های مقاله

- دیتاست عمومی ارائه نشده است
- جزئیات کامل معماری مدل‌ها در حد کلی بیان شده است
- بازتولید دقیق نتایج مقاله بدون دسترسی به دیتاست دشوار است

### محدودیت‌های prototype این پروژه

- داده آزمایش کوچک است
  - مدل‌های استفاده‌شده ساده هستند و fine-tuning انجام نشده
  - استخراج arguments دقیق و کامل نیست و صرفاً به صورت heuristic انجام می‌شود
- با این وجود، هدف prototype مطابق دستور استاد، نمایش عملی ایده اصلی مقاله بوده است.

---

## ۱۲) ایده‌های ادامه و بهبود آینده

برای توسعه این پروژه به یک سیستم واقعی و نزدیک‌تر به مقاله، می‌توان:

1. استفاده از دیتاست‌های استاندارد مثل ACE2005 یا MAVEN
2. استفاده از مدل‌های transformer fine-tuned برای event extraction
3. تعریف similarity پیشرفته‌تر برای arguments
4. استفاده از روش‌های clustering برای ادغام رویدادهای مشابه
5. اضافه کردن threshold برای حذف رویدادهای کم‌اعتماد
6. استفاده از مدل‌های LLM برای argument completion

---

## ۱۳) کاربرد عملی روش مقاله در سامانه‌های واقعی

روش مقاله می‌تواند در سامانه‌های عملی زیر استفاده شود:

- سامانه پایش اخبار امنیتی و حملات
- تحلیل روندهای سیاسی و تصمیم‌گیری دولتی
- تحلیل بازارهای مالی و استخراج رخدادهاى اقتصادى
- سیستم‌های خلاصه‌سازی خبری و timeline generation
- سیستم‌های fact-checking یا news verification



ویژگی Confidence Score کمک می‌کند سیستم بداند کدام رویدادها برای نمایش یا تصمیم‌سازی مناسب هستند و کدام رویدادها نیاز به بررسی انسانی دارند.

---

## نتیجه‌گیری نهایی

این مقاله یک روش ensemble برای استخراج رویداد از اخبار پیشنهاد می‌دهد که در آن خروجی مدل‌های مختلف با هم ترکیب می‌شوند و با استفاده از معیار Confidence Score، میزان اعتماد به رویدادهای استخراج‌شده تخمین زده می‌شود.

نوآوری اصلی مقاله در معرفی CS و الگوریتم iterative update برای بهبود اعتماد خروجی‌هاست. این رویکرد باعث افزایش دقت و کاهش خطا در استخراج رویداد می‌شود و قابلیت استفاده در سیستم‌های عملی را دارد.

Prototype پیاده‌سازی‌شده در این پروژه نیز نشان می‌دهد که ایده مقاله حتی در مقیاس کوچک قابل پیاده‌سازی است و می‌تواند به عنوان پایه‌ای برای توسعه سیستم‌های واقعی‌تر در آینده مورد استفاده قرار گیرد.

---