

Amirkabir University of Technology (Tehran Polytechnic)

بازنمایی معنایی متن با استفاده از مدل های زبانی و کاربرد آن در جستجوی معنایی اطلاعات SDI

سعيده ممتازى

دانشیار گروه هوش مصنوعی

دانشگاه صنعتی امیرکبیر

momtazi@aut.ac.ir

سازمان نقشەبردارى كشور بهمن ۱۴۰۳

- دانشیار گروه هوش مصنوعی دانشکده مهندسی کامپیوتر دانشگاه صنعتی امیر کبیر
 - مسئول آزمایشگاه پردازش زبان طبیعی دانشگاه صنعتی امیرکبیر
- رییس پژوهشکده هوش مصنوعی و سامانههای هوشمند دانشگاه صنعتی امیر کبیر

- کارشناسی نرمافزار و کارشناسی ارشد هوش مصنوعی از دانشگاه صنعتی شریف
 - دکترای هوش مصنوعی از دانشگاه زارلند آلمان
 - پسا دکترا هوش مصنوعی از موسسه HPI آلمان
 - پسا دکترا هوش مصنوعی از موسسه DIPF آلمان



فهرست مطالب

- مقدمهای بر مفاهیم واژگان در پردازش زبان
 - بازنمایی متن
 - توليد متن
 - مدل های زبانی بزرگ
 - جستجوی معنایی با بازنمایی متن

معنای واژگان

pepper, n.

Pronunciation: Brit./'pεpə/, U.S./'pεpə/Forms: OE peopor (rare), OE pipcer (transmission error), OE pipor, OE pipur (rareFrequency (in current use):

Etymology: A borrowing from Latin. Etymon: Latin piper.

< classical Latinpiper,a loanword < Indo-Aryan (as is ancient Greek $\pi \acute{\epsilon}\pi \epsilon\rho\iota$); compare Sai

I. The spice or the plant.

1.

a. A hot pungent spice derived from the prepared fruits (peppercorns) of the pepper plant, *Piper nigrum* (see sense 2a), used from early times to season food, either whole or ground to powder (often in association with salt). Also (locally, chiefly with distinguishing word): a similar spice derived from the fruits of certain other species of the genus *Piper*; the fruits themselves.

The ground spice from *Piper nigrum* comes in two forms, the more pungent *black pepper*, produced from black peppercorns, and the milder *white pepper*, produced from white peppercorns: see BLACK *adj.* and *n*. Special uses 5a, PEPPERCORN *n*. 1a, and WHITE *adj.* and *n*.¹ Special uses 7b(a).

2.

a. The plant *Piper nigrum* (family Piperaceae), a climbing shrub indigenous to South Asia and also cultivated elsewhere in the tropics, which has alternate stalked entire leaves, with pendulous spikes of small green flowers opposite the leaves, succeeded by small berries turning red when ripe. Also more widely: any plant of the genus *Piper* or the family Piperaceae.

b. Usu. with distinguishing word: any of numerous plants of other families having hot pungent fruits or leaves which resemble pepper (**1a**) in taste and in some cases are used as a substitute for it.

c. U.S. The California pepper tree, Schinus molle. Cf. PEPPER TREE n. 3.

3. Any of various forms of capsicum, esp. *Capsicum annuum* var.

annuum. Originally (chiefly with distinguishing C. annuum Longum group, with elongated taste, the source of cayenne, chilli powder, perennial C. frutescens, the source of Taba (more fully **sweet pepper**): any variety of group, with large, bell-shaped or apple-sha usually ripening to red, orange, or yellow a cooked as a vegetable. Also: the fruit of any Sweet peppers are often used in their green immedure st



معنای واژگان



Pronunciation: brn. /'pɛpə/, U.S. /'pɛpər/

Forms: OE **peopor** (*rare*), OE **pipcer** (transmission error), OE **pipor**, OE **pipur** (*rare* **Frequency (in current use):**

Etymology: A borrowing from Latin. **Etymon:** Latin *piper*.

< classical Latin piper, a loanword < Indo-Aryan (as is ancient Greek $\pi \dot{\epsilon} \pi \epsilon \rho \iota$); compare Sa

I. The spice or the plant.

1.

a. A hot pungent spice derived from the prepared fruits (peppercorns) of the pepper plant, *Piper nigrum* (see sense 2a), used from early times to season food, either whole or ground to powder (often in association with salt). Also (locally, chiefly with distinguishing word): a similar spice derived from the fruits of certain other species of the genus *Piper*; the fruits themselves.

The ground spice from *Piper nigrum* comes in two forms, the more pungent *black pepper*, produced from black peppercorns, and the milder *white pepper*, produced from white peppercorns: see BLACK *adj.* and *n*. Special uses 5a, PEPPERCORN *n*. 1a, and WHITE *adj.* and *n*.¹ Special uses 7b(a).

2.

a. The plant *Piper nigrum* (family Piperaceae), a climbing shrub indigenous to South Asia and also cultivated elsewhere in the tropics, which has alternate stalked entire leaves, with pendulous spikes of small green flowers opposite the leaves, succeeded by small berries turning red when ripe. Also more widely: any plant of the genus *Piper* or the family Piperaceae.

b. Usu. with distinguishing word: any of numerous plants of other families having hot pungent fruits or leaves which resemble pepper (**1a**) in taste and in some cases are used as a substitute for it.

c. U.S. The California pepper tree, Schinus molle. Cf. PEPPER TREE n. 3.

3. Any of various forms of capsicum, esp. *Capsicum annuum* var.

annuum. Originally (chiefly with distinguishin C. annuum Longum group, with elongated taste, the source of cayenne, chilli powder, perennial C. frutescens, the source of Taba (more fully **sweet pepper**): any variety of group, with large, bell-shaped or apple-sha usually ripening to red, orange, or yellow a cooked as a vegetable. Also: the fruit of any Sweet peppers are often used in their green immature st



معنای واژگان

pepper, n. lemma

Pronunciation: DIT. /'pɛpə/, U.S. /'pɛpər/

Forms: OE **peopor** (*rare*), OE **pipcer** (transmission error), OE **pipor**, OP **pipur** (*rare* **Frequency (in current use)**:

Etymology: A borrowing from Latin. **Etymon:** Latin *piper*. < classical Latin *piper*, a loanword < Indo-Aryan (as is ancient Greek πίπερι); compare Sai

I. The spice or the plant.

a. A hot pungent spice derived from the prepared fruits (peppercorns) of the pepper plant, *Piper nigrum* (see sense 2a), used from early times to season food, either whole or ground to powder (often in association with salt). Also (locally, chiefly with distinguishing word): a similar spice derived from the fruits of certain other species of the genus *Piper*; the fruits themselves.

The ground spice from *Piper nigrum* codes in two forms, the more pungent *black pepper*, produced from black peppercorns, and the milder *white pepper*, produced from white peppercorns: see BLACK *adj.* and *s.* Special uses 5a, PEPPERCORN *n.* 1a, and WHITE *adj.* and *n.*¹ Special uses 7b(a).

a. The plant *Piper nigrum* (family Piperaceae), a climbing shrub indigenous to South Asia and also cultivated elsewhere in the tropics, which has alternate stalked entire leaves, with pendulous spikes of small green flowers opposite the leaves, succeeded by small berries turning red when ripe. Also more widely: any plant of the genus *Piper* or the family Piperaceae.

b.Usu. with distinguishing word: any of numerous plants of other families having hot pungent fruits or leaves which resemble pepper (**1a**) in taste and in some cases are used as a substitute for it.

c. U.S. The California pepper tree, *Schinus molle*. Cf. PEPPER TREE *n*. 3.

3. Any of various forms of capsicum, esp. *Capsicum annuum* var.

annuum. Originally (chiefly with distinguishing C. annuum Longum group, with elongated taste, the source of cayenne, chilli powder, perennial C. frutescens, the source of Taba (more fully **sweet pepper**): any variety of group, with large, bell-shaped or apple-sha usually ripening to red, orange, or yellow a cooked as a vegetable. Also: the fruit of any Sweet peppers are often used in their green immature st

sense





معنای واژگان

definition

pepper, n. lemma

 Pronunciation: brit.
 /'pεpə/, U.S.
 /'pεpər/

 Forms: OE peopor (rare), OE pipcer (transmission error), OE pipor, OP pipur (rare

 Frequency (in current use):

Etymology: A borrowing from Latin. **Etymon:** Latin *piper*. < classical Latin *piper*, a loanword < Indo-Aryan (as is ancient Greek πίπερι); compare Sai

I. The spice or the plant.

a. A hot pungent spice derived from the prepared fruits (peppercorns) of the pepper plant, *Piper nigrum* (see sense 2a), used from early times to season food, either whole or ground to powder (often in association with salt). Also (locally, chiefly with distinguishing word): a similar spice derived from the fruits of certain other species of the genus *Piper*; the fruits themselves.

The ground spice from *Piper nigrum* comes in two forms, the more pungent *black pepper*, produced from black peppercorns, and the milder *white pepper*, produced from white peppercorns: see BLACK *adj.* and *p.* Special uses 5a, PEPPERCORN *n.* 1a, and WHITE *adj.* and *n.*¹ Special uses 7b(a).

a. The plant *Piper nigrum* (family Piperaceae), a climbing shrub indigenous to South Asia and also cultivated elsewhere in the tropics, which has alternate stalked entire leaves, with pendulous spikes of small green flowers opposite the leaves, succeeded by small berries turning red when ripe. Also more widely: any plant of the genus *Piper* or the family Piperaceae.

b.Usu. with distinguishing word: any of numerous plants of other families having hot pungent fruits or leaves which resemble pepper (**1a**) in taste and in some cases are used as a substitute for it.

c. U.S. The California pepper tree, *Schinus molle*. Cf. PEPPER TREE *n*. 3.

3. Any of various forms of capsicum, esp. *Capsicum annuum* var.

annuum. Originally (chiefly with distinguishing *C. annuum* Longum group, with elongated taste, the source of cayenne, chilli powder, perennial *C. frutescens*, the source of Taba (more fully **sweet pepper**): any variety of group, with large, bell-shaped or apple-sha usually ripening to red, orange, or yellow a cooked as a vegetable. Also: the fruit of any Sweet peppers are often used in their green immeture st

sense





مفهوم در زبان

• مفهوم هر واژه معنایی است که آن واژه دربردارد

ارتباط میان واژهها در واقع ارتباط میان معانی آنها است و نه خود واژه

تعاريف رايج در مفهوم زباني

- شرط لازم
- شرط کافی



تعاريف رايج در مفهوم زباني

بیان ویژگیها

• بیان بافت



تعاريف رايج در مفهوم زباني

- بیان ویژگیها
 - بيان بافت





تعاريف رايج در مفهوم زباني

Ludwig Wittgenstein (1889-1951)

"The meaning of a word is its use in the language"



تعاريف رايج در مفهوم زباني

Zellig Harris (1954):

"If A and B have almost identical environments we say that they are synonyms."



- نمایش هر کلمه با یک بردار
- بردارها بهعنوان تعبيه واژگان
 - بیان شباهت معنایی واژگان

فهرست مطالب

- مقدمهای بر مفاهیم واژگان در پردازش زبان
 - بازنمایی متن
 - توليد متن
 - مدلهای زبانی بزرگ
 - جستجوی معنایی با بازنمایی متن

بازنمایی کلمات با روش Word2Vec



اموزش مدل

- عدم نیاز به آموزش مدلها با داده برچسبدار
 - برگرفته از مدلهای زبانی عصبی
 - Bengio et al. (2003) •
 - Collobert et al. (2011) •



Continious skip-gram

بازنمایی متن با روش **BERT**





75% Spam

BERT

Class

Spam

Spam

Not Spam

25% Not Spam

1 - Semi-supervised training on large amounts of text (books, wikipedia..etc).

The model is trained on a certain task that enables it to grasp patterns in language. By the end of the training process, BERT has language-processing abilities capable of empowering many models we later need to build and train in a supervised way.



Supervised Learning Step

















(c) Question Answering Tasks: SQuAD v1.1 Class Label



Single Sentence

(b) Single Sentence Classification Tasks: SST-2, CoLA



(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

معمارى شبكههاى ترانسفورمر





معماری شبکههای ترانسفورمر



معمارى شبكههاى ترانسفورمر



Transformer



معمارى شبكههاى ترانسفورمر



فهرست مطالب

- مقدمهای بر مفاهیم واژگان در پردازش زبان
 - بازنمایی متن
 - توليد متن
 - مدلهای زبانی بزرگ
 - جستجوی معنایی با بازنمایی متن

Generative Pre-Training (GPT)



GPT-1

• معماري ١٢ لايه



GPT-2

- معماری ۴۸ لایه
 افزایش پارامترها از ۱۱۷ میلیون به ۱٫۵ میلیارد



Model Dimensionality: 768

Model Dimensionality: 1024

Model Dimensionality: 1280

Model Dimensionality: 1600

• معماري ۹۶ لايه



175 billion parameters, 10x more than any previous non-sparse language mode.

3(

GPT-4

• معماری ۱۲۰ لایه



GPT Series



سایر مدلها

- GPT-Neo (6.7B), GPT-J (6B), GPT-NeoX (20B)
- LLaMA, 65B
- Mistral and Mixtral: 7B

فهرست مطالب

- مقدمهای بر مفاهیم واژگان در پردازش زبان
 - بازنمایی متن
 - توليد متن
 - مدلهای زبانی بزرگ
 - جستجوی معنایی با بازنمایی متن

قدرت مدلهای زبانی بزرگ



API RESEARCH

ChatGPT: Optimizing Language Models for Dialogue

S OpenAI

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests. ChatGPT is a sibling model to InstructGPT, which is trained to follow an instruction in a prompt and provide a detailed response.



Text Representation - Saeedeh Momtazi
OpenAI

- تاسیس در سال ۲۰۱۵
- اعضاى اوليه هيئت مديره:
 - Sam Altman
 - Elon Musk •
- سرمایه گذاری اولیه توسط مایکروسافت • یک میلیارد دلار در سال ۲۰۱۹ • ۱۰ میلیارد دلار در سال ۲۰۲۳
- استفاده از منابع زیرساختی ابرکامپیوترهای مایکروسافت



داده متنی آموزش

Dataset	Quantity (tokens)	Weight in Training mix
Common Crawl	410 bilion	60%
WebText2	19 bilion	22%
Books1	12 bilion	8%
Books2	55 bilion	8%
Wikipedia	3 bilion	3%

سناريوي أموزش

Step 1

Collect demonstration data and train a supervised policy.

0

Explain reinforcement

learning to a 6 year old.

We give treats and

punishments to teach ...

BBB

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3.5 with supervised learning.

Step 2

Collect comparison data and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

to train our

reward model.



reward model using the PPO reinforcement learning algorithm. Explain reinforcement

0

learning to a 6 year old.

0 · C · A · B

B Explain rewards

O

We give treats and

punishments to feach

A

learning, the agent is...

C

In machine

learning.

A new prompt is sampled from the dataset.

Optimize a policy against the

Step 3

The PPO model is initialized from the supervised policy.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.

Write a story about otters. Once upon a time...

 \mathbf{r}_k

کاربردهای ChatGPT

- چتبات
- نگارش خلاقانه
- ترجمه ماشینی
 - تکمیل متن
 - خلاصەسازى
- پرسش و پاسخ
- تولید کد و مستندسازی
- تولید سناریوی تست در چرخه تولید محصول

سایر مدل های زبانی منطبق شده

GPT4	Gemini	PaLM 2	Llama 2	Vicuna	Claude 2	Falcon	МРТ	Mixtral 8*7B	Grok	StableLM
OpenAl	Google	Google	Meta	LMSYS Org	Anthropic	Technology Innovation Institute	Mosaic	Mistral Al	xAI	Stability Al

Text Representation - Saeedeh Momtazi

سایر مدل های زبانی منطبق شده

GPT4	Gemini	PaLM 2	Llama 2	Vicuna	Claude 2	Falcon	МРТ	Mixtral 8*7B	Grok	StableLM	deepseek	
OpenAl	Google	Google	Meta	LMSYS Org	Anthropic	Technology Innovation Institute	Mosaic	Mistral Al	xAI	Stability Al	deepseek	

فهرست مطالب

- مقدمهای بر مفاهیم واژگان در پردازش زبان
 - بازنمایی متن
 - توليد متن
 - مدلهای زبانی بزرگ
 - جستجوی معنایی با بازنمایی متن

جستجوى معنايي اطلاعات SDI



عدم تطابق واژگان



معماري كلى موتور جستجوى اطلاعات SDI



رویکردهای جستجو

- به صورت کلی در این سامانه جستوجو به دو نوع انجام شده است:
 - جستوجوی واژگانی:
- این نوع جستوجو به جستجوی سنتی مبتنی بر کلیدواژه اشاره دارد که در آن موتور جستجو عبارات پرسوجو را با عبارات موجود در اسناد مطابقت می دهد.
 - جستوجوی معنایی:
- این نوع جستوجو پیشرفته تر و هوشمندانه تر است که در آن موتور جستجو معنا و هدف پرسوجو و اسناد را درک می کند و نتایج را بر اساس ارتباط و کیفیت محتوا برمی گرداند. عملکرد جستجوی معنایی سیستم پیشنهادی لایه ای از ارزش را به تجربه کاربر اضافه می کند، زیرا به دنبال درک مقصود کاربر در پشت پرسوجو و ارائه مجموعه ای شخصی تر و مرتبط تر از نتایج است.

جستجو در سطح عبارت

• مدل اسبرت





جستجو در سطح عبارت

• مقایسه مدلهای اسبرت

Model Name	Performance Sentence Embeddings (14 Datasets) 🕕	Performance Semantic Search (6 Datasets) 🕕	†≓ Avg. Performance ③	Speed	Model Size 🕕
all-mpnet-base-v2	69.57	57.02	63.30	2800	420 MB
multi-qa-mpnet-base-dot-v1 🕕	66.76	57.60	62.18	2800	420 MB
all-distilroberta-v1	68.73	50.94	59.84	4000	290 MB
all-MiniLM-L12-v2	68.70	50.82	59.76	7500	120 MB
multi-qa-distilbert-cos-v1 🕕	65.98	52.83	59.41	4000	250 MB
all-MiniLM-L6-v2	68.06	49.54	58.80	14200	80 MB
multi-qa-MiniLM-L6-cos-v1 🕕	64.33	51.83	58.08	14200	80 MB
paraphrase-multilingual-mpnet-base-v2 🕕	65.83	41.68	53.75	2500	970 MB
paraphrase-albert-small-v2	64.46	40.04	52.25	5000	43 MB
paraphrase-multilingual-MiniLM-L12-v2 🕕	64.25	39.19	51.72	7500	420 MB
paraphrase-MiniLM-L3-v2 🕕	62.29	39.19	50.74	19000	61 MB
distiluse-base-multilingual-cased-v1 🕕	61.30	29.87	45.59	4000	480 MB
distiluse-base-multilingual-cased-v2 🕕	60.18	27.35	43.77	4000	480 MB

نمونه خروجى

جستجوى لايه
جنبر
تایچ جستجو برای " شکار"
لايه: يناهگاه حيات وحش
سارمان: اداره كل حفاظت محيط زيست
استيلا: 0.17742714
لايه: منطنه شکار ممنوع
سازمان: اداره كل حفاظت محيط زيست
استيلا: 0.15129378
لایه: مناطق گردشگری و بوم گردی
سازمان: اداره کل میراث فرهنگی و گردشگری
استيان: 0.11500208
لايه: اتر طبيعي ملي
سازمان: اداره کل حفاظت محیط زیست

کانی	جستجوى لايه م
	جنبو
	نتايج جستجو برای "شکار"
	سازمان: اداره کل حفاظت محیط زیست لایه: منطقه شکار معنوع

لابه: محدوده و حريم ميراث فرهنگی و طبيعی و آثار تاريخی

دَامتگاه سنتی
چېنېو (Construction)
چ جستچو برای "اقامتگاه سنتی"
ایه: مداملی گردشگری و بوم گردی
سازمان: اداره کل میراث فرهدگی و گردنسگری
متير: 0.27134547
رید: پداهگاه حیات و حش
سازسان: اداره کل حفاظت محیط زیست
0.23690055 : 3444
(یه: محدوده و حزیم میزرات فرهدگی و طبیعی و آثار تاریخی
سازمان: اداره کل میزان فرهنگی و گردندگری
متبر: 0.23680598
(یه: اتر طبیعی ملی
سازمان: اداره کل حفاظت معبط زیست
0.17727542 : مس رد:
ریم: محدوده سوابق صدون مجون بهره برداری ان اراضی مداطق چهارگانه محیطزیست
سازمان: اداره کل حفاظت محیط زیست

جستجوى لايه مكانى
اقامنگاه سنتی
جستجو
نتایج جستجو برای "اقامتگاه سنتی"
هيچ تنيجهاي پافت نشد.

یادگیری برای جستجو

- استفاده از بازخورد کاربران در جستجوهای پیشین
- آموزش مدل های تطابق معنایی بهمنظور بهبود کیفیت جستجو

52

یادگیری برای جستجو

• استفاده از بازخورد کاربران در جستجوهای پیشین

پهنهندی جستج نتایج جستجو برای "پهنهندی" یویستم: سازمان مدیریت و برنامه ریزی امتیاز: 0.582577
نتایج جستجو برای "پهنهبندی" لایه: پهنهبندی فرونشست سیستم: سازمان مدیریت و برنامه ریزی
لایه: پهنهبندی فرونشست سیستم: سازمان مدیریت و برنامه ریزی
سیستم: سازمان مدیریت و برنامه ریزی
امتياز: 0.582577
مدل داده: 1
 مرتبط) غیرمرتبط
لايه: بخش
سیستم: استانداری
امتياز: 0.44402462
مدل داده: 1
 مرتبط فیرمرتبط

یادگیری برای جستجو

• آموزش مدل های تطابق معنایی به منظور بهبود کیفیت جستجو



جمعبندى





جمعبندى













جمعبندى

































1×d

Embedding u

Mean pooling

Bert

Sentence A

Similarity

cosine-sim(u,v)

Embedding v

.

Mean pooling

Bert

Sentence B

 $X_{|V|}$

 $1 \times |V|$

Similarity

Sigmoid

Embedding z

Bert

Sentence B

Sentence A

 y_k

 $1 \times V$













سپاس فراوان از توجه شما