SparTerm:
Learning Term-based Sparse Representation for Fast Text Retrieval

Qun Liu (刘群)

Huawei Noah’s Ark Lab

智能信息检索与挖掘专题论坛
2021北京智源大会，2021-06-06

HUAWEI

SparTerm: Learning Term-based Sparse Representation for Fast Text Retrieval

Yang Bai*†
Tsinghua University

Xiaoguang Li*
Huawei Noah's Ark Lab

Gang Wang
Huawei Noah's Ark Lab

Chaoliang Zhang
Huawei Noah's Ark Lab

Lifeng Shang
Huawei Noah's Ark Lab

Jun Xu
Renmin University of China

Zhaowei Wang
Huawei Noah's Ark Lab

Fangshan Wang
Huawei Technologies Co., Ltd

Qun Liu
Huawei Noah's Ark Lab

ABSTRACT

Term-based sparse representations dominate the first-stage text retrieval in industrial applications, due to its advantage in efficiency, interpretability, and exact term matching. In this paper, we study the problem of transferring the deep knowledge of the pre-trained language model (PLM) to **Term**-based **Spar**se representations, aiming to improve the representation capacity of bag-of-words(BoW) method for semantic-level matching, while still keeping its advantages. Specifically, we propose a novel framework SparTerm to directly learn sparse text representations in the full vocabulary space. The proposed SparTerm comprises an importance predictor to predict the importance for each term in the vocabulary, and a gating controller to control the term activation. These two modules cooperatively ensure the sparsity and flexibility of the final text representation, which unifies the term-weighting and expansion in the same framework. Evaluated on MSMARCO dataset, SparTerm significantly outperforms traditional sparse methods and achieves state of the art ranking performance among all the PLM-based sparse models.

Query	Can hives be a sign of pregnancy?	
Type	Term frequency	SparTerm
Literal term Weights	hives are caused by allergic reactions . the dryness and stretching of your skin along with other changes can make you more susceptible to experiencing hives during pregnancy . hives can be caused by an allergic reaction to almost anything . some common causes of hives during pregnancy are noted below : medicine	hives are caused by allergic reactions . the dryness and stretching of your skin along with other changes can make you more susceptible to experiencing hives during pregnancy . hives can be caused by an allergic reaction to almost anything . some common causes of hives during pregnancy are noted below : medicine
Term expansion		symptoms:1.0, women:0.99, rash:0.98, feel:0.99, causing:0.97, body:0.96, affect:0.96, baby:0.94, pregnant:0.93, **sign:0.91**, ...

Figure 1: The comparison between BoW and SparTerm representation. The depth of the color represents the term weights, deeper is higher. Compared with BoW, SparTerm is able to figure out the semantically important terms and expand some terms not appearing in the passage but very semantically relevant, even the terms in the target query such as "sign".

Cite as: arXiv:2010.00768

Content

Introduction: Sparse vs. Dense Representation

Related Work: Neural Sparse Representation

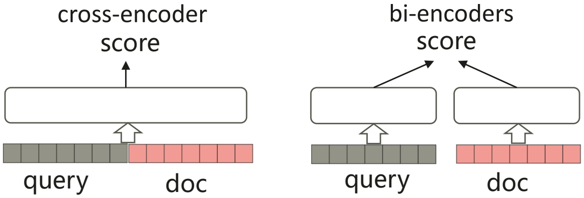
SparTerm

Conclusion and Future Work

Content

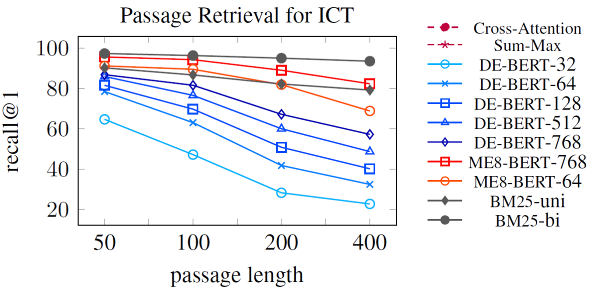
Some Preliminaries for Fast Text Retrieval

- Text Matching Paradigms
  - Cross-encoder
  - Bi-encoders
- Representation
  - Sparse High-dim Vector
  - Dense Low-dim Vector



Sparse or Dense Representation for Text Retrieval?

- For Exact Lexical Matching:
  - BM25 performs the best
  - Improve dense by increasing the dim and #vectors, but still worse than BM25

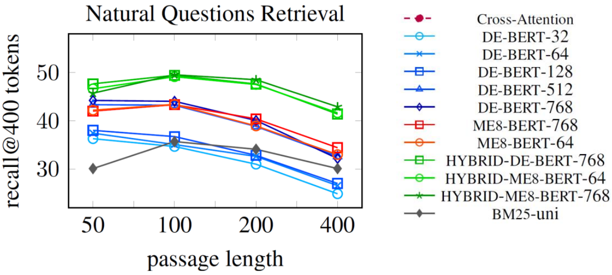


Query	She was built by the Harland and Wolff shipyard in Belfast.
Doc	RMS Titanic was the largest ship afloat at the time she entered service and was the second of three Olympic-class ocean liners operated by the White Star Line. She was built by the Harland and Wolff shipyard in Belfast. Thomas Andrews, chief naval architect of the shipyard at the time, died in the disaster.

Yuan et al., Sparse, Dense, and Attentional Representations for Text Retrieval, Google

Sparse or Dense Representation for Text Retrieval?

- For Semantic Matching:
  - BM25 performs the worst
  - PLM-based dense models show advantages to address the “lexical mismatch” problem



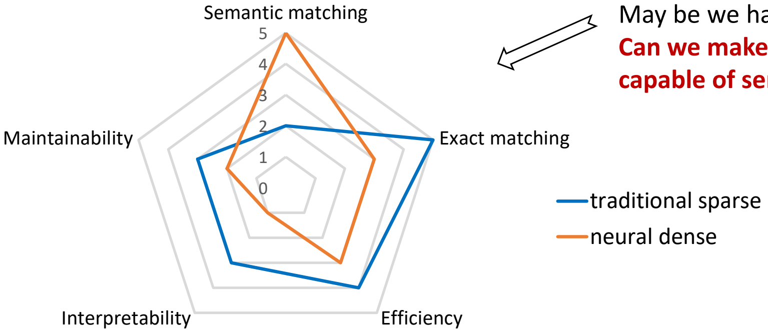
Query: Which city builds the Titanic ship?

Doc: RMS Titanic was the largest ship afloat at the time she entered service and was the second of three Olympic-class ocean liners operated by the White Star Line. She was built by the Harland and Wolff shipyard in Belfast. Thomas Andrews, chief naval architect of the shipyard at the time, died in the disaster.

Yuan et al., Sparse, Dense, and Attentional Representations for Text Retrieval, Google

Sparse or Dense Representation for Text Retrieval?

- Moreover, for industrial scenarios we have to consider:
  - Efficiency: Processing >50 billions docs
  - Interpretability: Predictable retrieval results
  - Maintainability: Easy to update



May be we have a hexagon like this, **Can we make sparse method also capable of semantic matching?**

HUAWEI

What Makes a Good Sparse Representation?

- Two aspects for improving sparse representation
  - Representation capacity: distinguishing ability for similar inputs
  - Representation sparsity: the proportion of # zero elements
- Improving representation capacity
  - For hot queries, we need better term weights
  - For rare queries, we need a "unbiased" words distribution estimation

Query: Medication for gum disease

Drugs Used to Treat Gum Disease Antibiotic treatments can be used either in combination with surgery and other therapies, or alone, to reduce or temporarily eliminate the bacteria associated with gum disease or suppress destruction of the tooth's attachment to the bone

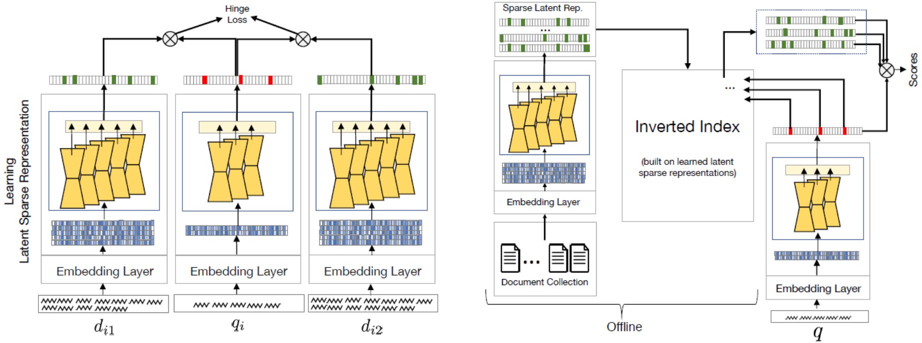
unobserved terms →

how, medication, doctors, medicine, cure, healing,...

Content

SNRM: Standalone Neural Ranking Model

- Learning sparse representation on latent space
  - Training optimized for information retrieval
  - Efficiently retrieve/inference using inverted index

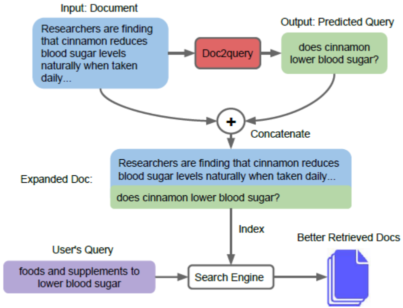


(a) Training time (b) Inference time

Zamani, Hamed, et al. From neural re-ranking to neural ranking: Learning a sparse representation for inverted indexing.

Doc2Query: Seq2seq term expansion

- Document Expansion by Query Prediction
  - Expanded terms bring better literal term weights
  - Expanded terms help narrow the “lexical mismatch” gap
  - T5 brings significant improvements over from-scratch model



	MRR@10 Dev	Test	R@1000 Dev	Latency (ms/query)
BM25 (Anserini)	0.184	0.186	0.853	55
doc2query, top-k, 10 samples	0.218	0.215	0.891	61
docTTTTTquery, top-k, 5 samples	0.259	-	0.929	58
docTTTTTquery, top-k, 10 samples	0.265	-	0.939	61
docTTTTTquery, top-k, 20 samples	0.272	-	0.944	62
docTTTTTquery, top-k, 40 samples	0.277	0.272	0.947	64
docTTTTTquery, top-k, 80 samples	0.278	-	0.945	66
DeepCT [2]	0.243	0.239	0.913	55
Best non-ensemble, non-BERT [5]	0.290	0.277	-	-
BM25 + BERT Large [7]	0.375	0.368	0.853	3,500

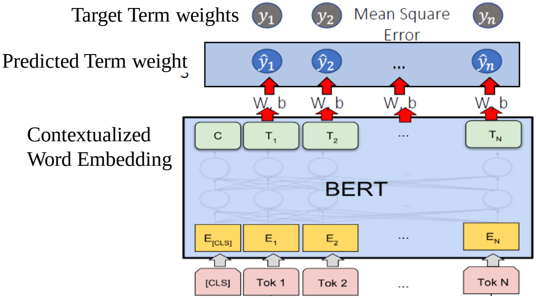
Table 1: Main results on MS MARCO the passage retrieval task.

Nogueira, Rodrigo, et al. Document expansion by query prediction.

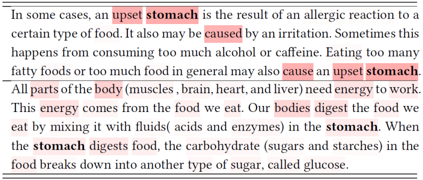
HUAWEI

DeepCT(HDCT): PLM-based term weighting

- Context-Aware Passage Term Importance Estimation
  - A term weights regression model based on PLM
  - Supervision of document term weights: relevant query, anchor text···



From frequencies to meanings!



In some cases, an upset **stomach** is the result of an allergic reaction to a certain type of food. It also may be caused by an irritation. Sometimes this happens from consuming too much alcohol or caffeine. Eating too many fatty foods or too much food in general may also cause an upset **stomach**. All parts of the body (muscles , brain, heart, and liver) need energy to work. This energy comes from the food we eat. Our bodies digest the food we eat by mixing it with fluids(acids and enzymes) in the **stomach**. When the **stomach** digests food, the carbohydrate (sugars and starches) in the food breaks down into another type of sugar, called glucose.

Dai, Zhuyun, and Jamie Callan. Context-Aware Term Weighting For First Stage Passage Retrieval.

HUAWEI

Content

Introduction: Sparse vs. Dense Representation

Related Work: Neural Sparse Representation

SparTerm

Conclusion and Future Work

Content

SparTerm
- Method
- Evaluation
- Analysis

Learning a term-based sparse representation in the full vocab space

- Better capacity: full vocabulary weighting
- Better sparsity/term activation: decoupled design of weighting and sparsification

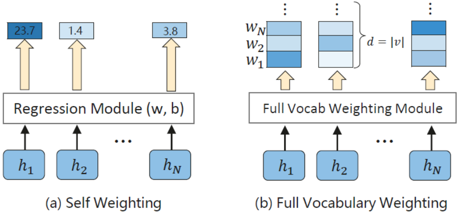
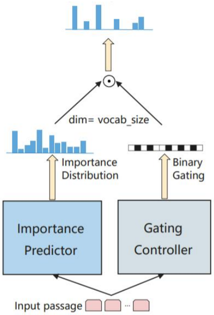


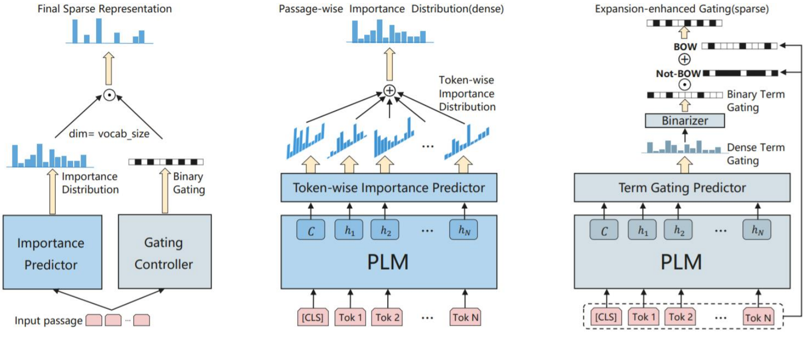
Figure 2: Full vocabulary weighting vs. self-weighting. In the self-weighting mechanism, each contextualized term representation only predicts the term weight for itself, while in the full vocabulary weighting, each term predicts a weight distribution in the full vocabulary.



HUAWEI

The model architecture

- The Importance Predictor: predict the importance for each term in the vocabulary
- The Gating Controller: control the term activation

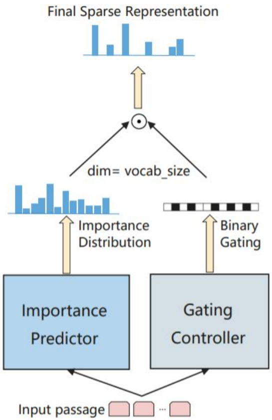


(a) SparTerm Model (b) Importance Predictor (c) Gating Controller

HUAWEI

Combination of importance predictor and gate controller

$$p' = \mathcal{I}(p) \odot \mathcal{G}(p)$$

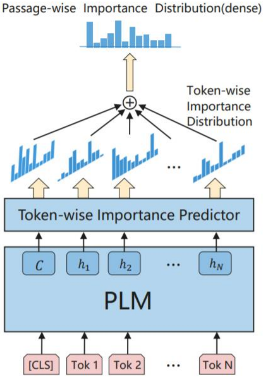


(a) SparTerm Model

Importance Predictor

$$I_i = LayerNorm(GeLU(h_i E_1))E_2^\top + b$$

$$I = \sum_{i=0}^{L} ReLU(I_i)$$



(b) Importance Predictor

HUAWEI

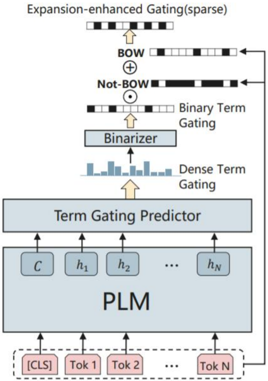
Gate Controller

$$G^{'} = Binarizer(G)$$

$$G_i^{'} = \begin{cases} 1, & \text{if } G_i > k \\ 0, & \text{if } G_i \leq k \end{cases}$$

$$G^e = G^{'} \odot (\neg BoW(p))$$

$$\mathcal{G}(p) = G^e + BoW(p)$$



(c) Gating Controller

HUAWEI

Training Object

$$L_{rank}(q_i, p_{i,+}, P_{i,-}) = -\log \frac{e^{\left\langle q_i^{'}, p_{i,+}^{'} \right\rangle}}{\sum_{j=1}^{M} e^{\left\langle q_i^{'}, p_{j,+}^{'} \right\rangle} + \sum_{k=1}^{N} e^{\left\langle q_i^{'}, p_{i,k,-}^{'} \right\rangle}}$$

$$L_{exp} = -\lambda_1 \sum_{j \in \{m | T_m = 0\}} log(1 - G_j) - \lambda_2 \sum_{k \in \{m | T_m = 1\}} logG_k$$

$$L = L_{rank} + L_{exp}$$

Training the Expansion-augmented Gating Controller

Table: Different kinds of term expansion.

Expansion type	Description and examples
Passage2query	Expand words that tend to appear in corresponding queries, e.g. “how far”.
Synonym	Expand synonym for original core words, e.g. “cartoon”->“animation”.
Co-occurrence	Expand co-occurrence words for original core words, e.g. “earthquakes”->“ruins”.
Summarization	Expand words that tend to appear in passage summarization or taggs.

Content

SparTerm

- Method
- Evaluation
- Analysis

Evaluation on public tasks

- MSMARCO(community QA task from Microsoft Bing Search)

Model	MS MARCO Passage Dev (full ranking)					MS MARCO Passage Local Eval (full ranking)					TREC2019DL Passage	MS MARCO Doc Dev
	MRR@10	R@10	R@50	R@100	R@1000	MRR@10	R@10	R@50	R@100	R@1000	NDCG@10	MRR@10
BM25*	19.47	40.59	61.47	69.33	85.71	18.68	38.68	58.25	66.18	85.94	50.61	24.52
Doc2query*	21.98	44.66	65.31	72.19	89.27	14.98	31.75	50.26	59.22	81.42	51.40	-
Doc2query-T5*	27.68	54.11	**75.61**	**81.89**	**94.71**	26.69	54.21	74.38	81.71	**94.66**	**64.20**	-
DeepCT*†	24.30	49.00	69.00	76.00	91.00	24.16	47.99	68.30	75.32	90.73	55.10	28.70
EPIC+BM25†	27.30	-	-	-	-	-	-	-	-	-	-	-
Dense Retrieval‡	30.80	-	-	-	92.80						59.40	-
SparTerm	**31.26**	**56.42**	75.29	81.60	93.80	**30.46**	**55.71**	**75.47**	**81.79**	93.84	59.32	**30.57**
SparTerm-literal	28.41	52.33	71.82	78.14	91.28	27.60	50.90	70.16	77.06	90.91	56.04	28.50

A comparable top ranking performance to PLM-based dense model!

HUAWEI

Evaluation on public tasks

- ICT(Extremely lexical matching) and NQ(Extremely semantic matching)

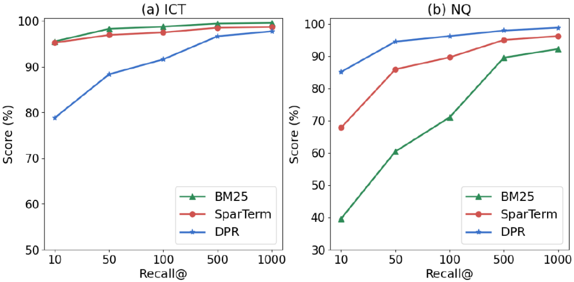


Figure 7: Performance of different models on ICT and NQ tasks. For both tasks, we use the recall of the golden passage to measure the performance.

Combining results of both tasks, SparTerm achieves a good balance between exact lexical matching and semantic-level matching!

HUAWEI

Evaluation on commercial datasets

- Auto-evaluation and human-evaluation for SparTerm on commercial scenarios

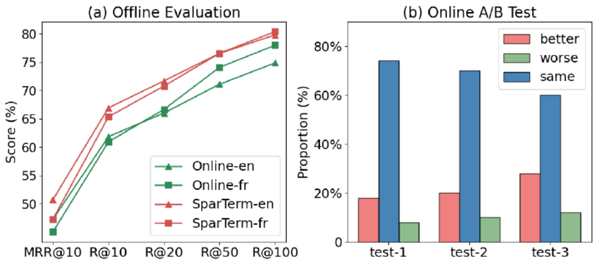


Figure 4: Performance of the product baseline and SparTerm on Commercial Dataset. (a) shows the results of automatic metrics. (b) shows the results of A/B test by human evaluation. "-en" denotes the English version and "-fr" the French version.

SparTerm improves end2end response relevance of search engine!

Content

Why SparTerm works?

- Performance under Various Lexical Overlaps

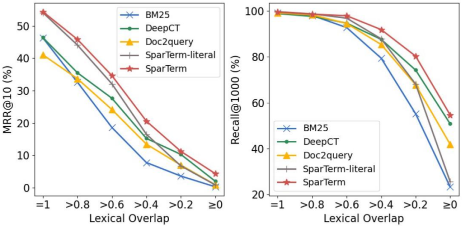


Figure 5: Performance changes of models under different lexical overlaps.

Table 4: The partition intervals of different overlap levels and the number of queries in each level. L represents the level of lexical overlap between the query and the ground truth passages. O is the range of overlap rate and N is the number of queries in each level.

L	=1	>0.8	>0.6	>0.4	>0.2	>0
O	o=1	0.8<o<1	0.6<o≤0.8	0.4<o≤0.6	0.2<o≤0.4	0≤o≤0.2
N	736	857	2630	1900	728	129

SparTerm woks on both hot queries and rare queries!

HUAWEI

Why SparTerm works?

- Performance under Various Lexical Overlaps

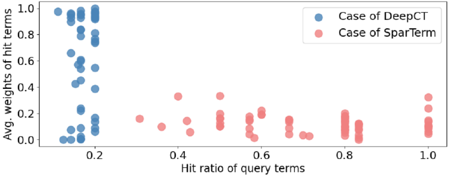


Figure 6: Query terms hit ratio and average weights (normalized) distribution for "good cases" of SparTerm and DeepcT. "good cases" refers to cases that the relevant document is ranked to top-1000 when $0 \leq o \leq 0.2$ in Table 4.

DeepCT obtains more sharp weight distribution and "put all bets" on the potentially most discriminate words.

SparTerm increases the lexical overlap by term expansion, therefore hitting more terms in queries to improve its retrieval performance

HUAWEI

Literal term weighting compared to DeepCT

- DeepCT obtains sparser and sharper distributions
- SparTerm “rewards” more words that are contextually-relevant and topic related

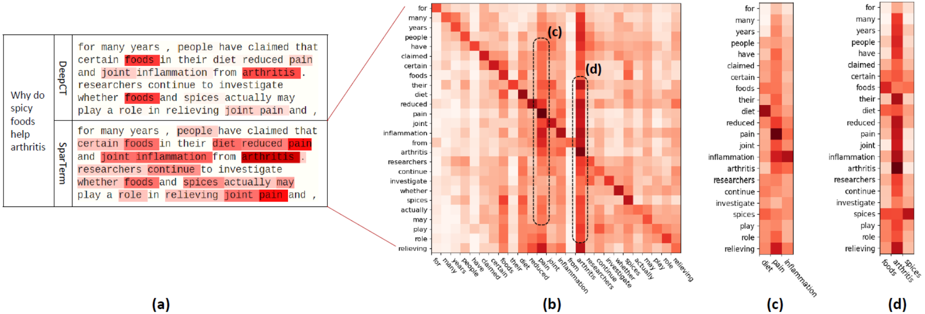


Figure 8: Term weightings of different passages weighted by DeepCT and SparTerm, and the mutual weighting contribution matrix predicted by SparTerm. The depth of the color represents the term weights, deeper is higher.

How terms are expanded?

- Passage2Query:

  $** \rightarrow$ how

- Synonyms:

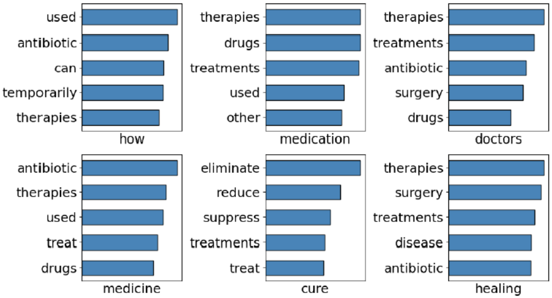
  drugs $\rightarrow$ medication

- Co-occurrence:

  season, heat $\rightarrow$ summer

Maybe the general knowledge from PLM?

Query:	Medication for gum disease
Passage:	Drugs Used to Treat Gum Disease Antibiotic treatments can be used either in combination with surgery and other therapies, or alone, to reduce or temporarily eliminate the bacteria associated with gum disease or suppress destruction of the tooth's attachment to the bone.
Expanded terms:	how, medication, doctors, medicine, cure, healing, ...



HUAWEI

Ablation studies verify

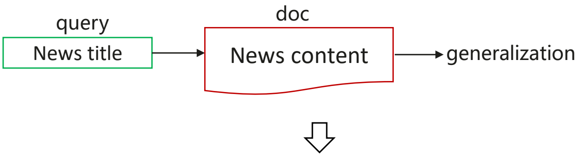
- The claim that term expansion can benefit from the e2e ranking optimization (#2)
- The necessity of retaining literal terms forcibly (#3)
- The effectiveness of our decoupling of term weighting and sparsification (#4,#5,#6)

#	Model	MRR@10	R@10	R@50	R@100	R@500	R@1000	Sparsity
1	SparTerm	31.26	56.42	75.29	81.60	91.19	93.80	99.46
2	SparTerm (w/o joint learning)	30.51	54.92	74.47	81.34	90.76	93.21	99.37
3	SparTerm (w/o retaining literal)	20.03	40.13	61.10	67.80	81.65	85.39	99.52
4	SparTerm (w/o gating controller,w/ topk-sparsification)	28.79	51.87	71.19	77.69	88.31	91.49	99.16
5	SparTerm (w/o gating controller,w/ th-sparsification)	30.12	53.74	73.47	79.73	89.5	92.57	0.00
6	SparTerm (w/o gating controller, w/ L1-sparsification)	22.98	41.47	62.05	69.81	82.24	86.44	98.65

HUAWEI

Further more, let SparTerm do more things that TF-IDF can do

- News tagging
- Key phrase extraction



Text

原 标题 ： 微软 ： 来 这个 开源 的 网站 看看 我们 是 如何 拥抱 开源 的 来源 ： 开源 中国 微软 近日 上
线 了 一个 新 的 开源 网站 — — 网站 本身 既 是 开源 的 ， 内容 也 是 关于 开源 的 — — 来 展示 其 如
何 拥抱 开源 ， 同时 提供 一些 开源 服务 。 从 首页 来看 ， 这一 开源 网站 的 核心理念 是 “ 开放 ” 、
“ 协作 ” 和 “ 灵活 ” 。 微软 在 网站 中 陈列 了 自己 的 开源 项目 和 服务 。 网站 分为 参与 、 项目
、 生态 、 招聘 及 博客 等 版块 。 其中 ， “ 参与 ” 页面 还会 实时 显示 微软 各个 githubrepo 的 最
新 动态 。 该 网站 由

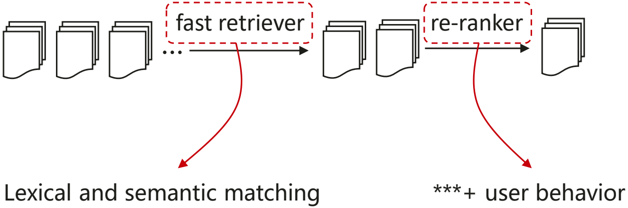
Label

微软：来这个开源的网站看看我们是如何拥抱开源的

Even the terms in titles/queries are limited(biased), SparTerm can recognized other important terms!

How SparTerm helps the commercial search engine

- SparTerm has been applied in the **first-stage retrieval** of $**$ search engine
- Significant improvements over online method on **Human Diff Evaluation**
- Indexing efficiency optimization: **20 billion doc titles/day**
- Support **multilingual** versions



This is what SparTerm focuses on!!

HUAWEI

Content

Introduction: Sparse vs. Dense Representation

Related Work: Neural Sparse Representation

SparTerm

Conclusion and Future Work

Conclusions and Future Work

- SparTerm: A term-based sparse representation learning method
  - A better trade-off of representation capacity vs. sparsity
  - A framework has been applied in commercial search engine
- Future Work
  - Large scale pre-training task for SparTerm, towards: stable, un-biased performance
  - Multi-grained term weighting
  - Combination of sparse and dense methods

Towards stable and un-biased term weighting

- Stably outperforms BM25 in all scenarios w/o specific fine-tuning?
  - Self-supervised training task designing(use BM25/query likelihood signals?)
  - Training on large user click data

Still challenging!

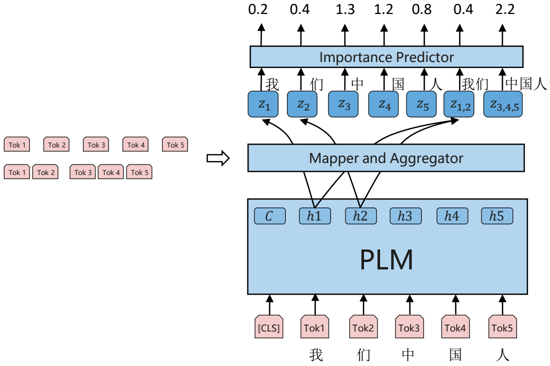
	BioASQ 7			BioASQ 8			Forum Travel			Forum Ubuntu		
	MAP	Prec @10	nDCG @10	MAP	Prec @10	nDCG @10	MAP	Prec @10	nDCG @10	MAP	Prec @10	nDCG @10
NEURAL MODELS												
ICT⋆	9.31*	3.84*	11.44*	9.31*	3.36*	11.78*	3.66*	11.60*	12.04*	8.93*	21.60*	23.21*
Ngram⋆	9.17*	3.86*	11.53*	8.81*	2.84*	10.74*	10.00	25.60	28.53	9.44*	22.00*	23.90*
QA†	17.80*	7.46*	21.93*	14.61*	4.26*	17.09*	11.00	27.60	28.32	17.78	**34.00**	34.73
QGen‡	**32.45**	**13.48**	**37.23**	**30.32**	**9.36**	**34.53**	**11.79**	**32.00**	**33.34**	**17.97**	32.40	**36.11**
TERM/HYBRID MODELS												
BM25⋆	45.12*	**20.66**	50.33*	38.61*	11.94*	42.78*	15.41*	37.60	39.21	16.23*	31.20*	35.16*
QGenHyb‡	**46.78**	20.60	**52.16**	**41.73**	**12.84**	**46.18**	**18.19**	**40.80**	**43.92**	**21.97**	**39.60**	**43.91**

Ma, Ji, et al. "Zero-shot Neural Passage Retrieval via Domain-targeted Synthetic Question Generation"

Multi-grained term weighting

- There is a tokenization gap between PrLMs and IR systems
  - PrLMs use wordpieces while IR systems use words and phrases
  - The tokenization(granularity) gap is bigger for Chinese

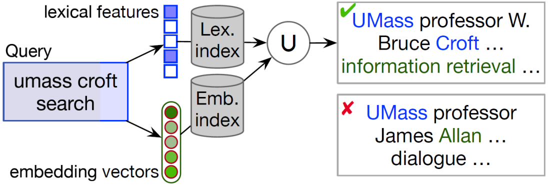
We may need to try multi-grained term weighting!



Combination of sparse and dense methods

- How to combine sparse and dense representation
  - Ensemble style (DUALRM by Gao, Luyu , et al.)
  - Sparse for Doc but dense for term? (COIL by Gao, Luyu , et al.)

$$s_{\text{DUALRM}}(q,d) = \lambda_{\text{test}} s_{\text{lex}}(q,d) + s_{\text{emb}}(q,d)$$



Gao, Luyu, et al. "Complement lexical retrieval model with semantic residual embeddings.".

HUAWEI

Content

Thank you!

把数字世界带入每个人、每个家庭、
每个组织，构建万物互联的智能世界。

Bring digital to every person, home and organization for a fully connected, intelligent world.

HUAWEI