CHASE: Commonsense-Enriched Advertising on Search Engine with Explicit Knowledge

Chao Zhang[1†], Jingbo Zhou[2†*], Xiaoling Zang[1], Qing Xu[1], Liang Yin[1],
Xiang He[1], Lin Liu[1], Haoyi Xiong[2], Dejing Dou[2]
[1]Baidu Search Ads (Phoenix Nest), Baidu Inc. [2]Baidu Research
{zhangchao38, zhoujingbo, zangxiaoling, xuqing06, yinliang01}@baidu.com
{hexiang, liulin03, xionghaoyi, doudejing}@baidu.com

ABSTRACT

While online advertising is one of the major sources of income for search engines, pumping up the incomes from business advertisements while ensuring the user experience becomes a challenging but emerging area. Designing high-quality advertisements with persuasive content has been proved as a way to increase revenues through improving the Click-Through Rate (CTR). However, it is difficult to scale up the design of high-quality ads, due to the lack of automation in creativity. In this paper, we present Commonsense-Enriched Advertisement on Search Engine (CHASE) — a system for the automatic generation of persuasive ads. CHASE adopts a specially designed language model that fuses the keywords, commonsense-related texts, and marketing contents to generate persuasive advertisements. Specifically, the language model has been pre-trained using massive contents of *explicit knowledge* and fine-tuned with well-constructed quasi-parallel corpora with effective control of the proportion of commonsense in the generated ads and fitness to the ads' keywords. The effectiveness of the proposed method CHASE has been verified by real-world web traffics for search and manual evaluation. In A/B tests, the advertisements generated by CHASE would bring 11.13% CTR improvement. The proposed model has been deployed to cover three advertisement domains (which are kid education, psychological counseling, and beauty e-commerce) at Baidu, the world's largest Chinese search engine, with adding revenue of about 1 million RMB (Chinese Yuan) per day.

CCS CONCEPTS

• **Information systems → Sponsored search advertising**; **Information extraction**; *Computational advertising*.

ACM Reference Format:
Chao Zhang[1†], Jingbo Zhou[2†*], Xiaoling Zang[1], Qing Xu[1], Liang Yin[1],, Xiang He[1], Lin Liu[1], Haoyi Xiong[2], Dejing Dou[2]. 2021. CHASE: Commonsense-Enriched Advertising on Search Engine with Explicit Knowledge. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM '21), November 1–5, 2021, Virtual Event, QLD, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3459637.3481902

1 INTRODUCTION

For the last decades, online advertising has been proved as one of the most successful business models and the major income source of internet industries [27]. The global online advertising market has grown four times in the last decade, especially in search ads[1]. The internet giants, such as Google, Amazon, and Facebook, earn hundreds of billions of USDs in their U.S. advertising revenue every year[2]. As the world's largest Chinese Search Engine, Baidu also keeps a record of tens of billions of RMB (Chinese Yuan) revenue quarterly in the Chinese online advertising market[3]. Given its fast-growing nature, the online advertising market has become a competitive field, where internet companies compete with each other to promote products, services, and ideas from advertisers to a large population of potential online customers through advertisement distribution [12].

From advertisers' perspectives, the major concern of distributing ads online is the effectiveness of advertising [18] against monetary costs for ad display [31]. One way to represent the effectiveness of advertising is to use Click-Through Rates (CTRs), while online advertising distributors all try to maximize the opportunities for ad displays (i.e., displaying ads in banners, or interrupting the video with clips for ads) and improve the CTRs through personalized recommendation with respect to browsing records of users [16, 36].

In addition to display and distribution, generating persuasive ads subject to users' needs is yet another solution for algorithms to boost the business performance of online advertising for major internet players [8, 29, 34, 39]. Existing approaches mainly focus on generating customized/contextual ads [34] to make content fit the user interfaces and contexts of web pages, or automatically designing relevant/personalized contents [8, 29, 39] that fit users' interests and intentions. While these efforts successfully adapt the contents of ads to users, they sometimes are not persuasive enough to encourage users to click through the ad or purchase the goods. Long-term studies in advertising find that incorporating knowledge, namely *persuasion knowledge* [15], in content of advertisement

[†]Equal contribution. [*]Corresponding author.

CIKM '21, November 1–5, 2021, Virtual Event, QLD, Australia
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8446-9/21/11…$15.00
https://doi.org/10.1145/3459637.3481902

[1] https://www.statista.com/statistics/276671/global-internet-advertising-expenditure-by-type/

[2] https://www.cnbc.com/2020/06/22/google-ad-revenue-will-drop-this-year-emarketer-says.html

[3] https://ir.baidu.com/index.php/news-releases/news-release-details/baidu-announces-third-quarter-2020-results

may be more attractive and persuasive to the users from cognitive perspectives [7, 23, 28].

Hence, our work intends to study the problem of persuasive ad generations using open common-knowledge sources. Specifically, given **(a)** marketing materials, including short description to the goods, slogans, and business information provided by advertisers, and **(b)** high-frequency contents in explicit knowledge bases, including free encyclopedia and discussion threads. The goal of our work is to automatically generate contents of ads that enrich advertisers' marketing materials using the explicit knowledge subject to the ads' keywords (also named bidwords) and title.

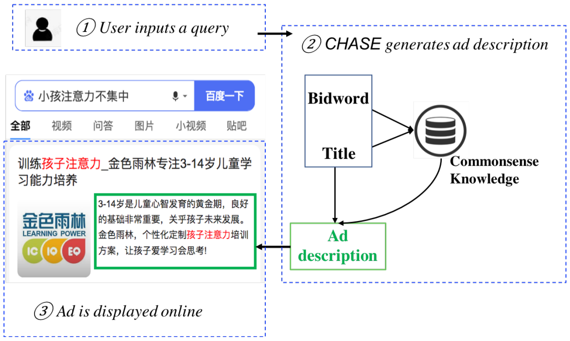
① User inputs a query
② CHASE generates ad description
小孩注意力不集中
百度一下
全部
视频
问答
图片
小视频
贴吧
训练孩子注意力_金色雨林专注3-14岁儿童学习能力培养
金色雨林
LEARNING POWER
IC
IQ
EQ
3-14岁是儿童心智发育的黄金期，良好的基础非常重要，关乎孩子未来发展。金色雨林，个性化定制孩子注意力培训方案，让孩子爱学习会思考!
Bidword
Title
Commonsense Knowledge
Ad description
③ Ad is displayed online

Figure 1: An illustration for the workflow of CHASE.

To this end, we present CHASE— the Commonsense-enricHed Advertising on Search Engine at Baidu that incorporates the persuasion knowledge extracted from the explicit knowledge bases for automatic ad generation. We illustrate a workflow of CHASE in Figure 1. The main function of CHASE is shown in step 2 of Figure 1. After user inputting a query (step 1 in Figure 1), CHASE will generate a persuasive advertisement description conditioned on the bidwords, advertisement title and external commonsense knowledge. After step 2, the generated ad description will be displayed on the website online (as shown in step 3 in in Figure 1).

We address the above discussed technical issues in the design of CHASE and make three unique contributions as follows. (1) In this work, we study the problem of knowledge-enriched ad generation for search engines. To the best of our knowledge, this work is the first to study the automatic generation of online ad with respect to advertiser-provided marketing materials, and explicit knowledge sources with persuasive information. (2) We design and implement CHASE with advanced novel language models for multiple source fusions. Specifically, we propose a novel method to construct a large scale of quasi-parallel corpora. A novel knowledge-guided generation and commonsense adapter mechanism are also investigated to generate human-readable and persuasive language presentation for advertising purposes. (3) We deploy CHASE for realistic ad generation and distribute the generated ads in three advertisement domains (which are kid education, psychological counseling, and beauty e-commerce) to the public through Baidu search. The A/B tests, in comparison with the state-of-the-art models, show that CHASE can improve the CTR by 11.13% as increasing the revenue of about 1 million RMB (Chinese Yuan) per day. The ablation studies further confirmed the effectiveness of all components of CHASE.

2 RELATED WORKS

In this section, we first review the backgrounds and related works, then discuss the most relevant works to our study.

2.1 Data-driven advertising

Data-driven techniques have been widely used for improving the quality of online advertising [14, 20, 30, 33, 35, 37, 43, 48]. In addition to matching well-drafted ads and the search queries, works have been done to generate the ads from keywords of search queries and marketing materials [1, 3, 8, 20, 29, 43, 44]. Keyword generation has also been widely adopted for ad matching [1, 49], information retrieval [4, 50], question answering [9, 45], and so on. The generation of complete ads has been proposed since the rise of natural language processing (NLP) techniques [3]. More recently, generating the long sequence of texts for advertising on search engines becomes possible with deep reinforcement learning algorithms [20], while template-driven techniques [43] still play a critical role for generating relevant contents subject to the search. Furthermore, some patent technologies [8, 29, 44] have been recently proposed to generate rich contents for advertising through incorporating various devices, mediums, and data sources. The effectiveness of online advertising techniques could be evaluated by CTRs under A/B tests [18]. Surveys on data-driven techniques for search, recommendation, and online advertising could be found in [36, 48].

2.2 Language models for text generation

In terms of methodologies, our work is also relevant to the efforts of text generation and deep generative language models. While general purpose language models [11, 41] have been proposed to perform various NLP tasks, as was discussed, certain fusion [25, 47], control [19, 26, 46] and adaptation [10, 42] techniques are required to improve the generation of ads.

Given the context of language and sources of knowledge, Zhao et al. [47] propose to use a generative language model to fuse the knowledge with contexts for language generation. Specifically, the model first embeds the language context and retrieved contents of knowledge into vectors, then encodes them into latent spaces using language and knowledge encoders respectively. Further, the proposed algorithm models the joint probability of a word using the context processor, document reader, and language model. The algorithm generates the knowledge-enriched texts through sequencing the words of maximal joint likelihoods. In addition to using contents retrieved, Koncel-Kedziorski et al. [25] adopt knowledge graphs for structured language generation, where the knowledge graph provides both structural control of the language and the knowledge as enrichment in the generated texts. Compared to [25, 47] that proposed to extract knowledge from either retrieved contents or knowledge graphs for fusion and generation, CHASE uses both retrieved contents and knowledge graphs for generation and control.

In order to control the structure and elements of generated texts, various control mechanisms have been proposed [19, 20, 25, 26, 46, 52]. Specifically, these works could be categorized as two types: (1) structural control and (2) latent variables/attributes manipulation. For structural control, the algorithms consider the generated texts as a sequence of vocabularies and the goal of control is to be with certain structures, such as graphs [25] or steps of control processes [19, 20]. For latent variables/attributes manipulation, these algorithms usually first map the generated texts in a latent space of semantics and syntax, then they control the generated texts through manipulating the variables [26, 46]. Both ways of language generation control rely on prior knowledge on the either structures or contents for better generation.

To fit the context of language (e.g., e-commerce, search by queries) for the text presentation, there frequently needs to adapt the text generation with respect to the contexts of language. Chen et al. [10] investigate ways to generate description of goods for e-commerce contexts, where the knowledge on the products and characteristics of customers has been used to personalize the generated contents according to the products and the interests of customers. Further, Wang et al. [42] propose to adapt the long texts subject to the short query in a sequence-to-sequence generation setting, where they incorporate the attentions of texts and queries for improved generation.

In terms of research problems, a relevant work to our study is Aiad [43], both of us intend to generate ads subject to queries and marketing materials for Baidu Search Engine. However, the main purpose of [43] is to generate an optimal combination of advertisement components (including buttons, images, titles) by a template-driven method which does not involve text generation and control.

In summary, CHASE made significant contributions compared to above works. Our later experiments based on real internet traffics with A/B tests and the extensive third-party manual evaluation would further confirm the advantages of CHASE for commonsense-enriched advertising on search engine.

3 CHASE: OVERALL SYSTEM DESIGN & IMPLEMENTATIONS

In this section, we provide an overall system design of CHASE. We first introduce the preliminaries overall this paper. Then we give an introduction about the offline advertisement generation (inference) process of CHASE. Finally, we briefly discuss the online ad matching process of Baidu.

3.1 Preliminaries

At first we introduce the basic setting and notations using throughout the paper. The application setting is a standard sponsored search. In this paper, we focus on the persuasive advertisement description generation problem, one of the most important inventions in Baidu's search advertisement system (well-known as the "Phoenix Nest" inside Baidu).

To set up an advertisement, advertisers first select k bidwords $b_i = \{b_{i,0}, b_{i,1}, ..., b_{i,k}\}$ related to their business. Then the advertiser also provides a multi-word title t_i and a multi-word description body d_i. Thus, each ad item ad_i is just a triple $ad_i =< b_i, t_i, d_i >$. Given a query q on a search engine, the search advertisement system first matches bidwords, and then retrieve the pre-designed advertisement. In real-life applications, the advertisement description body d_i usually is written by advertisers or automatically generated based on templates defined by advertisers or the sponsored search system [3].

The primary goal of CHASE is to generate persuasive ad description body d, which has a strong relation with the CTR which is the number of times an ad clicked divided by the number of times an ad displayed. CTR can be considered as a measure of "attractiveness" or "persuasiveness" of the ad body. CTR is directly related to profitability for search engines [13].

The key idea of CHASE is to bring more knowledge into the advertisement description body d to improve the CTR on the search engine. The effectiveness of this method can be explained from two perspectives. The first one is about the user search behavior. When a user begins to initiate a query on the search engine, she/he usually first wants to know some commonsense or background knowledge about the query, before trying to find a service or a product to solve their demand. Take a query about kid education as an example, given a user query of "How to do if children cannot focus in classroom", if the ad text provides some introductory content about the reason for child distractions, the user will have stronger intent to click the advertisement to seek for professional consulting and other services about kid education. The second one is because of the low quality of the ad description body. Actually, facing massive user intents in daily life, a majority of advertisers, especially the small and medium-sized clients, cannot afford to produce enough high-quality advertising materials. So that, the advertisers tend to use common marketing sentences and monotonous slogans, like "many free courses for you". However, users usually revolt against such straightforward marketing sentences, resulting in a poor reading experience and a low CTR. Our experimental evaluation also demonstrates that the commonsense-enriched advertisement with explicit knowledge can be more friendly and persuasive for users, leading to notable CTR improvement of the advertisement.

3.2 Offline persuasive ad generation

In this section, we introduce the advertisement generation process of CHASE after the model optimization. The training process is expounded in Section 4. As shown in Figure 2, given the title t_i and corresponding bidword set b_i of an ad, the goal is to generate an elaborately refined target description body for such advertisement. The knowledge (entities) about the product/service extracted from the original ad description and the focus points (extracted from the bidwords) are also input into the CHASE model as auxiliary inputs. Moreover, we further design a novel commonsense adapter mechanism that can control the relative ratio of commonsense knowledge and marketing content in the generated ad description with effective fusion.

In CHASE, we can model the advertisement description body generation as a context-aware commonsense-enriched dialogue response generation process. Given an advertisement item $ad_i =< b_i, t_i, d_i >$, the title t_i is considered as the first round of dialogue response to this query indicated by bidword set b_i. Then, the advertisement description body is considered as the second round of response where the bidword and the upper ad title can be considered as the context for the response in this round. Overall, CHASE takes $ad_i =< b_i, t_i, d_i >$ as context, and generates commonsense-enriched advertisement description body in the second round of response. Meanwhile, the commonsense adapter acts as a content controlling mechanism for the response generation, which makes

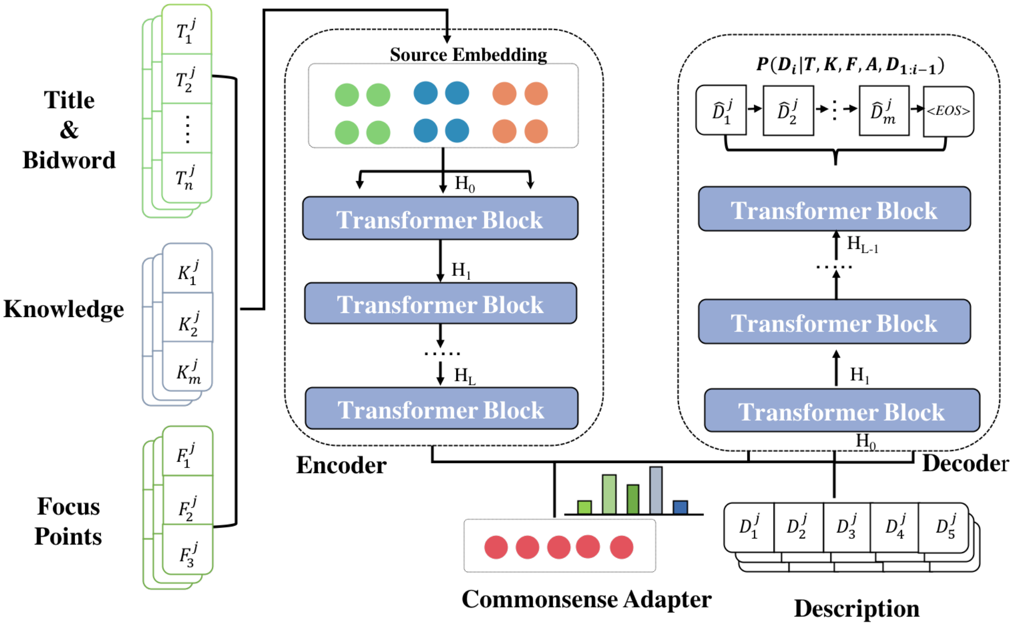

Figure 2: Illustration of the overall CHASE framework.

the advertisement description body not only contain the marketing information about the service/product, but also provide some commonsense knowledge about the service/product.

All the advertisement items are generated offline. CHASE processes billions of advertisements to refine their description body, and store them offline. The display of the advertisement is done by the online ad matching, which is introduced in the next section.

3.3 Online ad matching

Online ad matching mainly includes two steps on conventional sponsored search engines which are 1) advertisement retrieval and 2) advertisement ranking [13, 14, 21]. The advertisement retrieval step is to retrieve relevant ad candidates given a user query q. In this step, in order to retrieve all semantically relevant ad materials, many natural language processing (NLP) and query expansion technologies are employed [1, 2, 5, 51]. In ad ranking step, all candidates from the retrieval step are ranked according to several estimated business factors by machine learning models [14, 17] such as CTR and CVR (conversion rate). The top-ranked ads (usually 1-3 advertisements) are finally displayed on the search engine. A detailed description about the online ad matching on Baidu’s Search Advertisement system (a.k.a “Phoenix Nest”) can be seen in [14].

4 MODEL DESIGN

In this section, we first introduce how to construct the corpora for training CHASE. Then we briefly discuss how to pre-train CHASE with the masked sequence to sequence method. Finally, we give an in-depth discussion about the knowledge-guided generation as well as the commonsense adapter.

4.1 Constructing quasi-parallel corpora

A major challenge of CHASE is the lack of high quality parallel corpora. It is possible to manually rewrite a large scale of advertisement description body, and then to train an end-to-end encoder-decoder model to translate the original advertisement description body (which usually only contains marketing information) into a commonsense-enriched description (which contains some basic knowledge). However, such a method is not practical and almost impossible in real-life applications since the domain of advertisement is too complex, and the labor cost to annotate such parallel corpora is too high to be acceptable. Here we propose a novel strategy to automatically construct the quasi-parallel corpora with very low cost. In current implementation, the corpora are built on three advertisement domains which are kid education, psychological counseling and beauty e-commerce.

The general idea of the quasi-parallel corpora is that we construct a corpus which is a mixture of commonsense description corpora C^c and marketing description corpora C^m (from advertisement description). Then we use a knowledge-based filter method to reduce the data distribution difference between C^c and C^m. For an advertisement item $ad_i =< b_i, t_i, d_i >$, except its original linking to C^m, we also link some bidwords b_i to commonsense description C^c. Then we use bidword set b_i and title t_i as input, and alternatively use the C^c and C^m as output. In this way, the commonsense description corpora C^c and marketing text corpora C^m are indirectly linked by bidwords and titles. That is why we call our data as quasi-parallel corpora.

When to train CHASE with the quasi-parallel corpora, the model is partially optimized to generate the commonsense description and is partially optimized to generate the marketing description. Therefore, CHASE can be forced to learn to generate description with both commonsense description and marketing description. Moreover, we also introduce a commonsense adapter to control the ratio of the commonsense description overall advertisement description body, which is introduced in Section 4.4.

4.1.1 Corpora collection and advertisement synthesis. The commonsense description corpora are obtained from the following websites. Note that we only obtain the corpora related with three advertisement domains (i.e. kid education, psychological counseling and beauty e-commerce) in current implementation. Hereafter, we use C^c to conveniently refer to such commonsense description corpora. We will expand to cover as many domains as possible in future.

- **Baidu Baike**[4] is the world's largest online Chinese encyclopedia (just like Wikipedia in English). We use part of the Baidu Baike data to construct the commonsense description corpora C^c which is also a set of triples in the form $< b_i, t_i, d_i >$. Here the description of each encyclopedia entity is treated as d_i. The problem is how to construct 1) b_i and 2) t_i. For b_i, the name of the encyclopedia entity is included in the bidword set b_i directly. Moreover, we check the search behavior log on the Baidu search engine. The most frequent query about this encyclopedia entity in recent one month is also included in the bidword set b_i after word segment. For t_i, we use the co-click method to determine the title t_i. Before or after a user clicking the encyclopedia entity, the user may also click other webpages in a short time interval. We use the title of the most frequent co-click webpage as the title t_i. Only the encyclopedia entities in the domain of interest of Phoenix Nest system and having at least one click in recent one month are included in this corpora. In this way, we can construct millions of pseudo-advertisement triples.
- **Baidu Zhidao** [5] is the largest Chinese community-based question answering (CQA) site in the world. We also use Zhidao data to construct pseudo-advertisement triples $< b_i, t_i, d_i >$ of C^c. Here the question and answer of each QA item in Zhidao are treated as title t_i and description d_i respectively. We also use the search behavior log on the Baidu search engine to help to form the pseudo-advertisement triples. We use the most frequent query leading to click this QA item of Zhidao as the bidword set b_i (after word segment) in the recent one month. In this way we constructed millions of pseudo-advertisement triples from Zhidao data.
- **Article** We also crawled high quality articles from the web[6], and used this data to construct the pseudo-advertisement triples $< b_i, t_i, d_i >$. Similar to the Zhidao data, we treat the article title as title t_i, the article content as description d_i, and the most frequent query leading to click this article as the bidword set b_i (after word segment) in recent one month.

There are also a large scale of advertisements, which are provided by advertisers, in the Phoenix Nest system. The description body of this data mainly contains the marketing information. We use C^m to denote such advertisement data. We also use $C = C^c \cup C^m (ad_i =< b_i, t_i, d_i >\in C)$ to denote the whole corpora.

4.1.2 Knowledge-based selection. The data distribution between C^c and C^m is much different which hinders the model optimization using this data. To this end, we propose a knowledge-based selection method to relieve the distribution difference between C^c and C^m.

At first, we build a commonsense knowledge vocabulary V^c which is a set of words constructed from commonsense corpora C^c. For every sentence in the commonsense corpora $s^c_j \in C^c$, we first segment the sentence s^c_j with a word ranking toolkit.[7] All the words with importance < 2 are discarded.[8] After that, we count the occurrence of each leaf word and remove 1) the top-10% most frequent words and 2) the words with occurrence less than three. The reason to remove the most frequent words is that they do not represent the unique knowledge of commonsense knowledge since almost every document mentioned such words; and the reason to remove the low frequent word is to remove the noise word to avoid bringing errors.

For every advertisement triple $ad_i =< b_i, t_i, d_i >$, we define a commonsense ratio function $\lambda(\cdot)$ which can calculate the commonsense ratio which is the overlap between d_i and V^c divided by the length of d_i, i.e. $\lambda(d_i) = \frac{|d_i \cap V^c|}{|d_i|}$. For both C^c and C^m, we only keep item ad_i if it does not have too few or too much commonsense knowledge words, in other words, only the item with $\lambda^{down} \leq \lambda(d_i) \leq \lambda^{up}$ will be kept in the corpora.

The reason to remove the advertisement triples ad_i of both C^c and C^m with too large and too small commonsense ratio $\lambda(d_i)$ is to make the data distribution of C^c and C^m be similar. In other words, removing advertisement triples with $\lambda(d_i) < \lambda^{down}$ is because such items contain too few commonsense knowledge which cannot help the model to learn commonsense knowledge; and removing the items with $\lambda(d_i) > \lambda^{up}$ is because such triples cannot help the model to learn generate marketing related description.

4.2 Pre-trained language model

Before training the text generation model, we first use the quasi-parallel corpora to pre-train a language model to facilitate the downstream generation task. We adopt a masked sequence to sequence pre-training (MASS) for encoder-decoder based language generation [40]. Given an source sentence $s \in C$, we denote $s_{\overline{i:j}}$ as a sentence whose fragment from i to j of source sentence s is masked (the number of tokens being masked of s is $j - i + 1$). The optimization process of MASS is to pre-train a sequence to sequence auto-regressive encoder-decoder model by predicting the sentence fragment $s_{i:j}$ taking the masked sequence $s_{\overline{i:j}}$ as input. Formally, the log likelihood function can be expressed as:

$$L(\theta_{mass}; C) = \sum_{s \in C} logP_{mass}(s_{i:j}|s_{\overline{i:j}}, \theta_{mass}) \quad (1)$$

$$= \sum_{s \in C} log \prod_{k=i}^{j} P_{mass}(s_{k,\{u:v\}}|s_{<k,\{i:j\}}, s_{\overline{i:j}}, \theta_{mass}) \quad (2)$$

[4] https://baike.baidu.com/
[5] https://zhidao.baidu.com/
[6] For example, the article like https://baijiahao.baidu.com/s?id=1589801386751923407

[7] We use the lexical analysis tool for Chinese LAC(https://github.com/baidu/lac) [22], and other lexical analysis tools are also possible.
[8] LAC divide the words into four levels (0-3), and the word with importance <2 are stop words and redundancy words

Here P is the auto-regressive encoder-decoder framework with a stack of transformer blocks shown in Figure 2. An illustration of the pre-training process is shown in Figure 3.

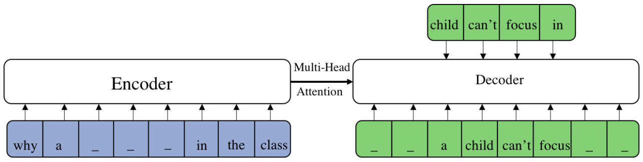

Figure 3: Illustration of masked sequence to sequence (MASS) pre-training.

4.3 Knowledge-guided generation

In this step, we introduce the knowledge-guided method to generate the advertisement description body. The key step is to extract the knowledge entities and focus points from the advertisement triples. Given an advertisement triple $< b_i, t_i, d_i >$, the objective of CHASE is to generate a better description d_i' that has both commonsense knowledge and marketing information.

As shown in Figure 2, we use a encoder-decoder framework to learn to refine the advertisement description body d_i. A straightforward method is to use the title and bidwords as the input of the encoder, and try to train the decoder to generate the description. However, there are two challenges for such a method. At first, the title and keywords have only limited information about the advertisement. The main topic of the description may be lost after the decoder. Second, for different intents of the advertisement, the description body is also quite diverse. For example, for the queries "how much is a kid educational class" and "what is a kid educational class", the description of the advertisement should be different. Therefore, instead of using the title and bidwords as the input of the encoder, we bring two auxiliary information for the encoders which are 1) entities from the advertisement description body and 2) the focus points from bidwords.

The entities are extracted from the advertisement description body after linking with a knowledge graph (KG) $\mathbb{G}$. We assume each description d_i contains n_i tokens that $d_i = \{x_{i,1}, x_{i,2}, ..., x_{i,n_i}\}$. The entity linking task is to map each token into an entity from the KG $\mathbb{G} = \{e_0, e_1, ..., e_{|\mathbb{G}|}\}$. Formally, the entity linking task can be formulated as:

$$G(d_i, \mathbb{G}) \xrightarrow{map} \{e_j\}_{j\in\{1,\cdots,|\mathbb{G}|\}}. \tag{3}$$

In CHASE, we use an enterprise entity linking API[9] to conduct the entity linking task.

We divide the advertisement triples into different classes according to their bidwords, which is called as "focus point". The major focus points are shown in Table 1. In our model, we use a set of pre-defined regular expression rules by considering the prefix or suffix of the bidwords to classify the focus points. The reason to use the rule-based method instead of building a classification model because:1) most of the bidwords can be matched by the regular expression rules to be classified, and 2) the rule-based method have high accuracy to avoid introduce error for the downstream generation task. Formally, given an intent set $\mathbb{I} = \{Price, Solution, Reason, Introduction, Enumeration\}$, we define a map function $I(\cdot)$ that can map the bidwords into one or several focus points:

$$I(b_i) \xrightarrow{map} \{I_j\}_{j\in\{1,\cdots,|\mathbb{I}|\}}. \tag{4}$$

After defining the function $G(\cdot)$ and $I(\cdot)$, the objective function to optimize the encoder-decoder framework $P_w(\cdot)$ can be formulated as:

$$\begin{aligned} L(\theta_w; C) &= \frac{1}{|C|} \sum_{ad_i \in C} logP_w(d_i|ad_i, \theta_w) & (5)\\ &= \frac{1}{|C|} \sum_{<b_i,t_i,d_i>\in C} logP_w(d_i|b_i, t_i, G(d_i, \mathbb{G}), I(b_i), \theta_w) & (6) \end{aligned}$$

Note that $P_w(\cdot)$ and $P_{mass}(\cdot)$ are the same model, and the optimization of $P_w(\cdot)$ is a fine-tuning of $P_{mass}(\cdot)$.

4.4 Commonsense adapter

CHASE is designed to generate advertisement descriptions that contain both commonsense information and marketing information. For this purpose, as shown in Figure 2, CHASE has a control mechanism, named as the commonsense adapter, to control the percent of commonsense knowledge in the whole advertisement description. We use λ^{ca} to denote the parameter of such a commonsense adapter. Given an advertisement document $ad_i =< q_i, t_i, d_i >$, the likelihood to generate an advertisement description d_i can be formulated as $P(d_i|ad_i, \lambda^{ca}, \theta_w)$. In real-life applications, setting the control parameter λ^{ca} into different levels (like 1 for 20%, 2 for 40%, ...) is good enough without requiring the control parameter λ^{ca} as a real number. Thus, to facilitate the optimization of the auto-regressive encoder-decoder $P(\cdot)$, we set the λ^{ca} as an integer with limited range that $\lambda^{ca} \in \mathbb{N}$ and $1 \le \lambda^{ca} \le \Lambda$. Then the objective function of Eqn. 5 to optimize the auto-regressive encoder-decoder framework with commonsense adapter can be reformulated as:

$$L(\theta_w; A) = \frac{1}{|C|} \sum_{ad_i\in C} \frac{1}{\Lambda} \sum_{\lambda^{ca}\in\Lambda} logP_w(d_i|ad_i, \lambda^{ca}, \theta_w) \tag{7}$$

During the training processing, we assume that $P_w(d_i|ad_i, \lambda^{ca}, \theta_w) = 0$ if $\lambda^{ca} \ne \lfloor \lambda(d_i) \rfloor$. Note that the commonsense adapter and knowledge-guided generation are simultaneously trained.

5 EXPERIMENTS

In this section, we conduct comprehensive experiments to evaluate the effectiveness of our model.

5.1 Experiment settings

5.1.1 Dataset and parameter settings. How to construct the dataset is already discussed in Section 4.1. For the evaluation purpose, we random select a dataset with 500,000 advertisement triples as test data, 800,000 advertisement triples as validation data. The size of training data (number of advertisement triples) is 10,800,000. During the experiment evaluation, if without specification, we set commonsense adapter parameter $\lambda^{ca} = 70\%$. For the knowledge-based selection, we set $\lambda^{down} = 20\%$ and $\lambda^{up} = 90\%$.

[9] http://kg.baidu.com/operatordetail/entityannotation

Focus Points	Description	Example (Chinese Example)
Price	Ask for the price for services or products	How much are kid education courses? (儿童早教课多少钱?)
Solution	Ask for the method to solve a problem	How to help a child focus in the classroom? (怎样帮助儿童上课集中注意力?)
Reason	Ask for the reason of a phenomenon	Why a child cannot focus in the classroom? (儿童上课注意力不集中是怎么回事？)
Introduction	Ask for the introduction of things	Is kid education important for children? (儿童早教是否重要？)
Selection	Select a list of items from a collection	Which type of kid education course is better? (哪种儿童早教课比较好?)

Table 1: The defined five query intents and their examples.

5.1.2 Setup. We implement all the models in PaddlePaddle[10]. The encoder and decoder have 6 blocks. The number of attention heads, embedding dimension and inner-layer dimension are 8, 256 and 512, respectively. The vocabulary dictionary is shared across all datasets and has size of 100k. Overall, the total number of model parameters is 85*1e6 (to be specific, it is 84, 969, 472).

Models are optimized with the Adam algorithm [24] using the learning rate of 1e-5, linear warmup over the first 2000 steps, and learning rate is polynomial decay. We run experiments on two V100 GPUs with maxtoken of approximately 10k. CHASE takes 2.5 hours to train 1 epoch. We run 30 epochs to pre-train the model by MASS, and run 5 epochs to fine-tune the model for knowledge-guided generation and commonsense adapter. During decoding, CHASE use beam search with beam size of 3, and CHASE does not allow to have repeated 3-gram.

5.1.3 Evaluation Metrics. To automatically evaluate the model performance, we use Perplexity (PPL) and Pairwise-BLEU to measure the model quality and diversity of generation results.

Perplexity [6] is an evaluation metric to measure the model capacity for language modeling which is the normalized inverse probability of the dataset. Formally, the perplexity metric (PPL) can be expressed as:

$$PPL(D) = \sqrt[n]{\frac{1}{P(d_1, d_2, ..., d_n)}} \tag{8}$$

where d_i is the target advertisement description body.

BLEU [32] is to use n-gram word matching to measure corpus similarity. In this paper, we use *Pairwise-BLEU* [38] to measure the diversity of generation results. Given a source sentence s_0, and $\{\hat{d}^1, ..., \hat{d}^K\}$ are K hypotheses. To measure similarity among the hypotheses, we compare them with each other. The more diverse the hypothesis set is, the lower the Pairwise-BLEU is. Formally, Pairwise-BLEU can be expressed as:

$$Pairwise - BLEU = BLEU([\hat{d}^i]; \hat{d}^j)$$
$$i \in \{1, ..., K\}; j \in \{1, ..., K\}; i \neq j$$

5.1.4 Baselines. We compare our model with following baselines to verify the effectiveness of our approach

- **CHASE:** is the proposed model in this paper. CHASE has been deployed online on Baidu search advertisement system (a.k.a Phoenix Nest) to cover three advertisement domains which are kid education, psychological counseling and beauty e-commerce.
- **PN-CS:** is the online deployed model on Phoenix Nest (before deploying CHASE) for automatic advertisement description generation. It is an encoder-decoder framework with six layers transformer. It is trained by high CTR advertisement data directly with bidwords+title as input and advertisement description as output.
- **PN-CP:** whose model architecture is the same with PN-CS, but PN-CP is pre-trained by commonsense corpora with MASS, and then is fine-tuned by advertisement data with high CTR. The input of the model is bidwords+title, and the output is also advertisement description.
- **PN-CH:** whose model architecture is the same as PN-CS. It is pre-trained by commonsense corpora with MASS, and is fine-tuned by marketing corpora after knowledge-based selection (see Section 4.1.2).
- **OAD**: is the original advertisement description written by advertisers.

Model	Perplexity ↓	Pairwise-BLEU ↓
PN-CS	30.95	85.03
PN-CP	24.36	79.71
PN-CH	15.91	72.18
CHASE	**8.28**	**47.89**

Table 2: Automatic evaluation of perplexity and pairwise-BLEU on test dataset. Bold scores are the best overall.

5.2 Automatic evaluation

Table 2 shows the automatic evaluation results. As we can see from Table 2, CHASE achieves the best performance among all baselines in terms of perplexity and pairwise-BLEU. For perplexity, a lower perplexity indicates the model can better predict the test sample by the language model. The reason is that we introduce the knowledge-guided generation method to provide more signal to control the quality of the generated results. The pairwise-BLEU of CHASE is better than all baselines which means the generation results of CHASE have high diversity. The reasons are:1) we build a quasi-parallel corpus to make CHASE generate advertisement description containing both commonsense knowledge and marketing information, and 2) we introduce a commonsense adapter mechanism to control the percent of commonsense knowledge which can avoid one kind of information to dominate the generated content.

[10] https://github.com/PaddlePaddle/Paddle

Model	Perplexity↓	pairwise-BLEU↓
CHASE	**8.28**	**47.89**
·w/o knowledge	9.13	50.49
·w/o focus point	9.26	59.82
·w/o comm-adapter	9.94	64.57

Table 3: Results of ablation study. Bold scores are the best.

Model	Readability↑	Coherence↑	Information↑	Overall↑
OAD	1.49	1.54	1.58	1.29
PN-CS	1.66	1.63	1.52	1.36
PN-CP	1.68	1.72	1.65	1.45
PN-CH	1.71	1.76	1.75	1.49
CHASE	**1.90**	**1.82**	**1.91**	**1.71**

Table 4: Manual evaluation results among baselines. Bold scores are the best.

Table 2 also shows that the pairwise-BLEU and perplexity of PN-CP are much better than PN-CS, which demonstrates the effectiveness of using commonsense corpora. All the automatic evaluation metrics of PN-CN are much better than PN-CN and PN-CP, which verifies that using knowledge-based selection can remove the low quality and monotone language corpora to improve the model performance. Note that we do not compare CHASE with OAD since advertisers only write one description for each advertisement which cannot be used to compute the perplexity and pairwise-BLUE. To sum up, experimental results of automatic evaluation show the superiority of CHASE.

5.2.1 Ablation study. To understand the impacts of each component of CHASE, we carry out ablation study by removing corresponding components as shown in Table 3. If without knowledge, all the metrics become worse which proves that the knowledge entities can empower the capacity of decoders. If without focus points, all the metrics become worse, especially for the pairwise-BLEU. This showcases that focus point can effectively guide the model to generate diverse advertisement description. If without the commonsense adapter, all the metrics become worse, especially for pairwise-BLEU. This demonstrates that commonsense adapter can affect the diversity of the description by controlling the ratio among commonsense knowledge and marketing information.

5.3 Manual evaluation

We conduct a manual evaluation on 300 random samples from our test dataset. Ten participants were recruited to measure the quality of the ad description generated by each baseline from four perspectives. Each perspective is measured by a 3-point Likert question where 0 is bad, 1 is neutral and 2 is good.

- **Readability**, which measures how the generated description is smooth and grammatically corrects.
- **Coherence**, which measure whether the description is relative with title and whether the generated result is consistent with the background knowledge.
- **Information**, which measures how informative the description is.
- **Overall**, measures the overall quality of the description.

As shown in Table 4, CHASE outperforms all baselines. For example, CHASE outperforms OAD (i.e. advertiser) by +0.42 score in overall quality, outperforms PN-CS (i.e. current online system on Phoenix Nest) by +0.35 score in overall quality. Moreover, CHASE also achieves higher scores than all baselines in readability and information obviously. The improvement indicates that CHASE can generate readable and informative advertisement descriptions for users after bringing commonsense knowledge into the model. From Table 4, we can find that the generated description of CHASE is also more coherent than other baselines. The coherence score also demonstrates that the knowledge-guided generation method and commonsense adapter can successfully control the consistency of generated results.

Since PN-CS is trained with the advertisement description with high CTR, it has good quality but tends to be monotonic. Thus, as we can see from Table 4, the readability and coherence of PN-CS are better than OAD, but the information of PN-CS is worse than OAD. Both PN-CP and PN-CH are better than PN-CS which demonstrates the usefulness of constructing commonsense corpora and the knowledge-based selection strategy.

Note that the CHASE can achieve higher score than OAD in all manual evaluation metrics. The reason is that the advertisers always buy many bidwords, but they do not have enough time to write detailed and creative advertisement descriptions for each bidword. Therefore, the advertisers usually use some common marketing description to fill such content. From user perspective, such advertisement description lacks enough information, is without good readability and has little coherence with the bidwords.

5.4 Online A/B test

We also conduct an online A/B test in three domains to show the superiority of CHASE. We used 5% real-world web traffics on Baidu search engine from three domains (which are kid education, psychological counseling, and beauty e-commerce) to conduct the A/B test. We had already got the permission from advertiser to use CHASE to generate advertisement description for this test. This online A/B test lasted for one week as shown in Table 6. In each day there were about 1 millions page views (with ad shows) for the testing. We use CTR compared with OAD to show the improvement of CHASE, which is defined as $\Delta CTR = \frac{CTR\ of\ model}{CTR\ of\ OAD}$. Except the displayed description, we keep other settings the same.

As we can see from Table 6, it is obvious that the CHASE significantly outperforms the baseline PN-CS. The overall ΔCTR of CHASE/OAD is higher than the one of PN-CS/OAD by 11.13%. The result clearly shows that the commonsense-enriched advertisements generated by CHASE are more persuasive and more likely to attract users to click on the advertisement. This CTR improvement of CHASE can increase the revenue of Phoenix Nest about 1 million RMB (Chinese dollars) per day.

Bidword: better, kid education, which	**Title**: Which better for kid education? normal teaching institute, powerful teaching resources!	
-	Original	For Better kid education? focus on kid education, professional enlightenment course for your children, lower cost, sign up now for free nice gift!
-	Online model (PN-CS)	Better kid education？ professional kid education, professional teaching team, let kids fall in love with learning.
Knowledge: kid education#enlightenment **Focus Point**: selection	CHASE (λ^{ca} = 40%)	Looking for better kid education？ enlightenment education is so essential for kids, necessary to choose carefully, professional teaching team, promote completely!
	CHASE (λ^{ca} = 70%)	Looking for better kid education？ child’s early stage before 6-year-old, called golden time, providing strong bases for further development and lifelong learning, deserving serious consideration and selection, high quality and professional teaching for curiosity arousing!

Table 5: Case Study.

Model Comparison	2021/1/18	2021/1/19	2021/1/20	2021/1/21	2021/1/22	2021/1/23	2021/1/24	Overall Average
PN-CS/OAD	+Δ 7.83%	+Δ 8.05%	+Δ 7.60%	+Δ 7.84%	+Δ 7.78%	+Δ 7.85%	+Δ 8.16%	+Δ 7.87%
CHASE/OAD	+Δ 18.82%	+Δ 18.94%	+Δ 19.01%	+Δ 19.06%	+Δ 19.05%	+Δ 19.03%	+Δ 19.08%	+Δ 19.00%

Table 6: Online A/B Testing of ΔCTR in a week. There are about 1 millions page views (with ad shows) per day for the testing.

Model Comparison	kid	psyc	beauty
PN-CS/OAD	+Δ 7.32%	+Δ 7.05%	+Δ 8.93%
CHASE/OAD	+Δ 19.92%	+Δ 21.65%	+Δ 17.13%

Table 7: CTR improvements under different domains. Kid represents kid education, psyc represents psychological counseling and beauty represents beauty e-commerce. During the testing, in each day there are about 0.68 millions, 0.13 millions and 0.27 millions page views (with ad shows) for kid, psyc and beauty respectively.

We also analyze the CTR improvements under different domains in Table 7. An interesting result is that the CTR improvement of CHASE in kid education domain and psychological counseling domain is larger than beauty e-commerce domain. The reason is that the commonsense knowledge requirement of kid education and psychological counseling is larger than beauty e-commerce. Besides, the advertisement description of beauty products is always more plentiful and fascinating compared with the other two domains. But the commonsense-enriched advertisement on beauty e-commerce can still significantly improve the CTR.

5.5 Case study

In this section, we perform a case study to observe how CHASE generated advertisement description. We present a real case selected from our system logs in Table 5. We show the bidword and title in the first row of Table 5. The knowledge and focus point are extracted from the original advertisement description written by advertisers. The right column of second row contains different advertisement descriptions generated by different methods. As we can see from Table 5, the original description written by advertisers only contain marketing information. Trained by the high CTR advertisement data, the online model PN-CS have learned to re-write the advertisement description by slightly reducing the marketing voice and add some friendly language like “let kids all in love with learning”. CHASE presents much different advertisement description. It first presents the commonsense knowledge about kid education, and then smoothly delivers some marketing information which is more delightful for users than the description of PN-CS. Besides, when λ^{ca} = 70%, CHASE generates the advertisement description with more background information than λ^{ca} = 40%, thus, the commonsense adapter parameter λ^{ca} can control the percent of the commonsense knowledge content.

6 CONCLUSIONS

In this work, we present CHASE— a real-world system deployed at Baidu Search Engine for automatic generation of persuasive ads with explicit knowledge. Given the marketing materials provided by the advertisers, CHASE generates advertisement description and enrich the texts using relevant commonsense knowledge. Specifically, CHASE adopts a novel language model that fuses language resources from marketing materials, and other commonsense knowledge contents for text generation, control and adaptation. The effectiveness of CHASE has been verified using real-world web traffics for search under A/B tests and third-party manual evaluation. In A/B tests, the advertisements generated by CHASE would bring 11.13% CTR improvement. The system and models have been deployed to cover three advertisement domains on Baidu, the world’s largest Chinese search engine, with a revenue growth about 1 million RMB (Chinese Yuan) per day. In future, we are planning to extend CHASE to support the advertisement in other domains which have high requirement for commonsense knowledge, such as automobile, real estate and adult education.

REFERENCES

[1] Vibhanshu Abhishek and Kartik Hosanagar. 2007. Keyword generation for search engine advertising using semantic similarity between terms. In *ICEC*. 89–94.

[2] Xiao Bai, Erik Ordentlich, Yuanyuan Zhang, Andy Feng, Adwait Ratnaparkhi, Reena Somvanshi, and Aldi Tjahjadi. 2018. Scalable query n-gram embedding for improving matching and relevance in sponsored search. In *KDD*. 52–61.

[3] Kevin Bartz, Cory Barr, and Adil Aijaz. 2008. Natural language generation for sponsored-search advertisements. In *EC*. 1–9.

[4] David B Bracewell, Fuji Ren, and Shingo Kuriowa. 2005. Multilingual single document keyword extraction for information retrieval. In *NLPKE*. 517–522.

[5] Andrei Broder, Peter Ciccolo, Evgeniy Gabrilovich, Vanja Josifovski, Donald Metzler, Lance Riedel, and Jeffrey Yuan. 2009. Online expansion of rare queries for sponsored search. In *WWW*. 511–520.

[6] Peter F Brown, Stephen A Della Pietra, Vincent J Della Pietra, Jennifer C Lai, and Robert L Mercer. 1992. An estimate of an upper bound for the entropy of English. *Computational Linguistics* 18, 1 (1992), 31–40.

[7] Raymond R Burke, Arvind Rangaswamy, Jerry Wind, and Jehoshua Eliashberg. 1990. A knowledge-based system for advertising design. *Marketing Science* 9, 3 (1990), 212–229.

[8] Ric Calvillo, Claude Denton, Joshua Allen Breckman, Per Anders Sandell, Derek J Yimoyines, Amit Deepak Adur, Christopher Connors, and Jonathan Palmer. 2020. System for high volume data analytic integration and channel-independent advertisement generation. US Patent 10,599,313.

[9] Jun Chen, Jingbo Zhou, Zhenhui Shi, Bin Fan, and Chengliang Luo. 2019. Knowledge abstraction matching for medical question answering. In *BIBM*. 342–347.

[10] Qibin Chen, Junyang Lin, Yichang Zhang, Hongxia Yang, Jingren Zhou, and Jie Tang. 2019. Towards knowledge-based personalized product description generation in e-commerce. In *KDD*. 3040–3050.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*. 4171–4186.

[12] Ulrich Dolata. 2017. *Apple, Amazon, Google, Facebook, Microsoft: Market concentration-competition-innovation strategies.* Technical Report. SOI Discussion Paper.

[13] Daniel C Fain and Jan O Pedersen. 2006. Sponsored search: A brief history. *Bulletin of the american Society for Information Science and technology* 32, 2 (2006), 12–13.

[14] Miao Fan, Jiacheng Guo, Shuai Zhu, Shuo Miao, Mingming Sun, and Ping Li. 2019. MOBIUS: towards the next generation of query-ad matching in baidu's sponsored search. In *KDD*. 2509–2517.

[15] Marian Friestad and Peter Wright. 1994. The persuasion knowledge model: How people cope with persuasion attempts. *Journal of consumer research* 21, 1 (1994), 1–31.

[16] Dinesh Gopinath and Michael Strickman. 2011. Personalized advertising and recommendation. US Patent App. 12/871,416.

[17] Thore Graepel, Joaquin Quiñonero Candela, Thomas Borchert, and Ralf Herbrich. 2010. Web-scale Bayesian click-through rate prediction for sponsored search advertising in Microsoft's bing search engine. In *ICML*. 13–20.

[18] Saikat Guha, Bin Cheng, and Paul Francis. 2010. Challenges in measuring online advertising systems. In *IMC*. 81–87.

[19] Xinting Huang, Jianzhong Qi, Yu Sun, and Rui Zhang. 2020. Generalizable and Explainable Dialogue Generation via Explicit Action Learning. In *EMNLP: Findings*. 3981–3991.

[20] J Weston Hughes, Keng-hao Chang, and Ruofei Zhang. 2019. Generating better search engine text advertisements with deep reinforcement learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2269–2277.

[21] Bernard J Jansen and Tracy Mullen. 2008. Sponsored search: an overview of the concept, history, and technology. *International Journal of Electronic Business* 6, 2 (2008), 114–131.

[22] Zhenyu Jiao, Shuqi Sun, and Ke Sun. 2018. Chinese Lexical Analysis with Deep Bi-GRU-CRF Network. *arXiv preprint arXiv:1807.01882* (2018). https://arxiv.org/abs/1807.01882

[23] Taylor Jing Wen, Eunice Kim, Linwan Wu, and Naa Amponsah Dodoo. 2020. Activating persuasion knowledge in native advertising: the influence of cognitive load and disclosure language. *International Journal of Advertising* 39, 1 (2020), 74–93.

[24] Diederik P Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

[25] Rik Koncel-Kedziorski, Dhanush Bekal, Yi Luan, Mirella Lapata, and Hannaneh Hajishirzi. [n.d.]. Text Generation from Knowledge Graphs with Graph Transformers. In *NAACL, Volume 1*. 2284–2293.

[26] Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2018. Multiple-attribute text rewriting. In *ICLR*.

[27] Mary Meeker and Liang Wu. 2018. Internet Trends 2018.

[28] Kotsedi D Monyeki, Han CG Kemper, Lateef O Amusa, and Marcus Motshwane. 2013. Advertisement and knowledge of tobacco products among Ellisras rural children aged 11 to 18 years: Ellisras Longitudinal study. *BMC pediatrics* 13, 1 (2013), 1–7.

[29] Ashwin Navin, Omar Zennadi, and David Harrison. 2020. Relevant advertisement generation based on a user operating a client device communicatively coupled with a networked media device. US Patent 10,567,823.

[30] Junwei Pan, Yizhi Mao, Alfonso Lobos Ruiz, Yu Sun, and Aaron Flores. 2019. Predicting different types of conversions with multi-task learning in online advertising. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2689–2697.

[31] Panagiotis Papadopoulos, Nicolas Kourtellis, and Evangelos P Markatos. 2018. The cost of digital advertisement: Comparing user and advertiser views. In *WWW*. 1479–1489.

[32] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*. 311–318.

[33] Claudia Perlich, Brian Dalessandro, Rod Hook, Ori Stitelman, Troy Raeder, and Foster Provost. 2012. Bid optimizing and inventory scoring in targeted online advertising. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. 804–812.

[34] Alec Reitter, Barb Chang, Ken Sun, Raghav Gupta, Alvaro Bolivar, and Alan Lewis. 2010. System and method for application programming interfaces for keyword extraction and contextual advertisement generation. US Patent 7,831,586.

[35] Kan Ren, Yuchen Fang, Weinan Zhang, Shuhao Liu, Jiajun Li, Ya Zhang, Yong Yu, and Jun Wang. 2018. Learning multi-touch conversion attribution with dual-attention mechanisms for online advertising. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 1433–1442.

[36] Maad Shatnawi and Nader Mohamed. 2012. Statistical techniques for online personalized advertising: A survey. In *SAC*. 680–687.

[37] Dou Shen, Arun C Surendran, and Ying Li. 2008. Report on the second kdd workshop on data mining for advertising. *ACM SIGKDD Explorations Newsletter* 10, 2 (2008), 47–50.

[38] Tianxiao Shen, Myle Ott, Michael Auli, and Marc'Aurelio Ranzato. 2019. Mixture models for diverse machine translation: Tricks of the trade. In *ICML*. 5719–5728.

[39] Sheridan Martin Small, Andrew Fuller, Avi Bar-Zeev, and Kathryn Stone Perez. 2012. Automatic Customized Advertisement Generation System. US Patent App. 12/886,141.

[40] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: Masked Sequence to Sequence Pre-training for Language Generation. In *ICML*. PMLR, 5926–5936.

[41] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223* (2019).

[42] Xinyi Wang, Jason Weston, Michael Auli, and Yacine Jernite. 2019. Improving conditioning in context-aware sequence to sequence models. *arXiv preprint arXiv:1911.09728* (2019).

[43] Xiao Yang, Daren Sun, Ruiwei Zhu, Tao Deng, Zhi Guo, Zongyao Ding, Shouke Qin, and Yanfeng Zhu. 2019. Aiads: Automated and intelligent advertising system for sponsored search. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1881–1890.

[44] Eric Zavesky, David Crawford Gibbon, Bernard S Renger, and Behzad Shahraray. 2020. Advertisement generation based on a user image. US Patent App. 17/020,024.

[45] Jie Zhao, Ziyu Guan, and Huan Sun. 2019. Riker: Mining rich keyword representations for interpretable product question answering. In *KDD*. 1389–1398.

[46] Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning Discourse-level Diversity for Neural Dialog Models using Conditional Variational Autoencoders. In *ACL*. 654–664.

[47] Xueliang Zhao, Wei Wu, Chongyang Tao, Can Xu, Dongyan Zhao, and Rui Yan. 2019. Low-Resource Knowledge-Grounded Dialogue Generation. In *ICLR*.

[48] Xiangyu Zhao, Long Xia, Jiliang Tang, and Dawei Yin. 2019. Deep reinforcement learning for search, recommendation, and online advertising: a survey. *ACM SIGWEB Newsletter* Spring (2019), 1–15.

[49] Hao Zhou, Minlie Huang, Yishun Mao, Changlei Zhu, Peng Shu, and Xiaoyan Zhu. 2019. Domain-constrained advertising keyword generation. In *WWW*. 2448–2459.

[50] Jingbo Zhou, Shan Gou, Renjun Hu, Dongxiang Zhang, Jin Xu, Airong Jiang, Ying Li, and Hui Xiong. 2019. A collaborative learning framework to tag refinement for points of interest. In *KDD*. 1752–1761.

[51] Jingbo Zhou, Qi Guo, HV Jagadish, Lubos Krcal, Siyuan Liu, Wenhao Luan, Anthony KH Tung, Yueji Yang, and Yuxin Zheng. 2018. A generic inverted index framework for similarity search on the gpu. In *ICDE*. 893–904.

[52] Meng Zhou, Jingbo Zhou, Yanjie Fu, Zhaochun Ren, Xiaoli Wang, and Hui Xiong. 2021. Description Generation for Points of Interest. In *ICDE*. 2213–2218.