3M3-5

検索連動型広告におけるテキスト自動生成とその評価指標の検討

Automatic Generation of Title and Description Texts for Sponsored Search Ads

馬場 惇 *1 Jun Baba　岩崎 祐貴 *1 Yuki Iwazaki　杉尾 樹 *1 Itsuki Sugio　北出 庫介 *1 Kosuke Kitade　福嶋 剛 *1 Takeshi Fukushima

*1株式会社サイバーエージェント
CyberAgent, Inc.

Sponsored search ads require short and expressive Titles and Descriptions to attract the attention of search engine users. Manually creating such Title and Description text is greatly time-consuming and difficult for Ad System operators. This paper presents a method to automatically generate Titles and Descriptions for sponsored search ads based on sentence network analysis, sentence contraction and auto summarization techniques. We evaluate the results of the text generation process using famous KPIs in ad system, such as click-through rate (CTR) and show that the proposed method achieves the same level of performance with human in the Title and Description text generation task.

1. はじめに

検索連動型広告は, 検索エンジンが検索クエリに応じて検索結果画面に表示させる広告であり, 広告がクリックされることで課金が発生するため広告掲載の費用対効果に優れていて, 検索エンジンマーケティング市場の大半の売上を占めている.

検索連動型広告は図 1 のようにテキスト形式であることが一般的で, 15 文字の Title 部と 38 文字の Description 部からなる. Title 部は見出しの役割を果たし, その詳細な説明を Description 部に記述する. 以下, Title 部と Description 部を合わせて TD と呼ぶ.

Google AdWords などが展開する検索連動型広告サービスでは, 広告の入札価格と, 広告表示回数に対するクリック回数の割合 (Click-Through-Rate: CTR) が高い TD が上位に掲載されるような戦略がとられている. 上記のような戦略をとる広告サービス上では, 検索エンジンを利用する一般ユーザの目を引きやすい TD を, ユーザが見飽きないように多種類用意することが重要になってくる. 現状の広告運用の現場では, そのような TD を人手で作成しているため, 広告運用コストが膨大になっており, 運用が追いつかないという問題が生じている.

図 1: Google AdWords の検索連動型広告の例

そこで本稿では, スマートフォンアプリの広告主を対象とし, TD 作成の工数を削減し運用コストを下げることを目標に, 平均的な効果のある広告テキストを, その広告に関連するテキストコンテンツから大量に, かつ, 自動的に生成する手法を提案する. 提案手法は 2 段階で構成されており, 入力として広告主のサービスや商品に関するテキストコンテンツ, 具体的にはスマートフォンアプリのアプリストアの説明文を利用する. まず第一段階では, 基となるテキストコンテンツから Description の候補となる文章を抽出する. そして次の段階で, 抽出された文章から Title の候補となる文字列を抽出する. 最後に, 各段階で抽出された Description 候補と Title 候補を組み合わせて,TD の候補リストとして利用する. 生成された候補リストを利用することで, 広告入稿者の運用コストを削減することができる. 本稿では, その運用コスト削減率や, 提案手法によって生成された TD の有効率, 広告効果を見るための CTR などの観点から提案手法を検証する.

本稿の構成は以下の通りである. 第 2 章では, 広告テキストの自動生成に関する先行研究について紹介し, 本稿の位置付けを明確にする. 第 3 章で提案するシステムについて全体像とそれを構成する手法について説明し, 第 4 章において提案手法の効果検証について紹介し, 第 5 章で提案手法の改善点について考察を述べる.

2. 関連研究

検索連動型広告のテキスト自動生成の取り組みとして, 幾島らの HTML 文書からの自動生成 [幾島 08] や, 藤田らの飲食店のサイトプロフィール文書からの自動生成 [藤田 11] が挙げられる. 両手法は本稿と同様に, 広告主のサービスや商品に関連するテキストコンテンツをもとに TD を生成する手法を採用している. 主に,tf-idf によるスコア関数によってテキストコンテンツからスコアの高い単語や文章を抽出し, 文節のスコアによって係り受け構造を加工して利用するアプローチで, それぞれ適用対象の文書に合わせたチューニングがなされている.

本稿では, スマートフォンアプリのストア説明文を対象テキストコンテンツとするため, 一般的でない未知語や新語, カタカナ語などの単語が多く含まれていたり, アプリの内容だけでなく, キャンペーン情報や対象端末, 注意点など, 本来広告として訴求したい内容以外の情報が多く含まれている. そのため, 文節の関係や係り受け構造が抽出しにくいという特性があり, 既存研究と同様の手法で文章の圧縮をすることが非常に難し

連絡先: 馬場 惇, 株式会社サイバーエージェント, Email: baba_jun@cyberagent.co.jp

い．また，ストア説明文は一度作成された後はあまり変更されない文章であることが多いため，多種類の TD を生成するためには工夫が必要となる．

そこで本稿では，一部の一般的でない単語に左右されずにアプリ内容を表す文章を抽出するために文章間の単語の共起に注目した重要文抽出手法を利用し，また,TD の種類を拡張するために，関連語や類義語によって文章内の単語を入れ替える拡張手法を用いる．文章圧縮の点においても，係り受け構造などを用いず，制約下での部分文字列抽出と体言止め表現への変換を行うことで，アプリストア説明文に対応できる TD 自動生成システムを構築する．

3. 提案手法

本稿で提案する広告テキスト自動生成システムの概要を図 2 に示す．

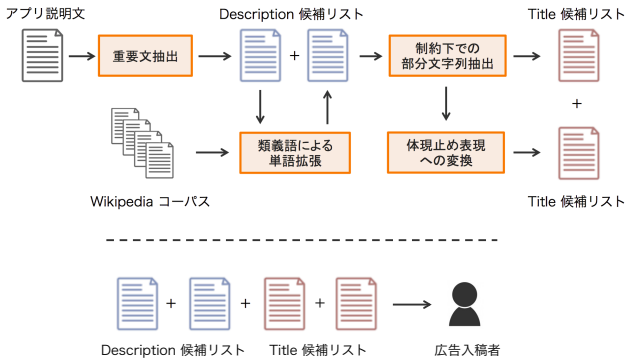

図 2: システム概要図

提案システムへの入力は，アプリストアに掲載している対象スマートフォンアプリの説明文である．提案システムは，入力テキストからの Description 候補抽出と，Description 候補からの Title 生成の 2 段階によって構成される．Description 候補抽出フェーズでは，アプリ説明文から重要文を抽出し，それらの文章の単語を類義語で入れ替え拡張を行うことで，アプリの内容を表し，かつ，十分な種類の Description の候補リストを出力する．Title 生成フェーズでは，出力された Description 候補を基に，制約に基づいた部分文字列を抽出したものを Title 候補として出力し，さらに，一定数以上の Title を確保するために，抽出された部分文字列のうち動詞句で終わるものを体言止め表現へと変形して,Title 候補に加える．最終的には，以上の処理によって出力された Title 候補と Description 候補を広告入稿者に提出し，そこから入稿者が TD として選択したものを配信する．以下の節で各フェーズにおける処理の詳細を述べる．

3.1 Description 候補の抽出

3.1.1 重要文抽出

入力コーパスとなるスマートフォンアプリの説明文中から，アプリ内容について述べた重要な文章を抽出してくるために，本研究では，テキスト自動要約の分野で提案された Salton らの文章ネットワークを構築する手法 [Salton 96] を応用する．Salton らは，ある文書において他の文章と単語の共起が多く見られる文章は代表的な文章である，という仮定に基づいている．各文章をノードに見立て，文章間の単語の共起から類似度を計算し，閾値以上の文章間をエッジでつないだ文章ネットワークから重要文章を抽出している．本稿ではこの手法を応用し，文章ネットワークを利用して重要文を抽出する．

具体的な例として以下のような 2 つ文章があったとする．

1. 簡単協力バトルで仲間を増やそう
2. ギルドバトルで仲間と勝利を掴もう

まず，各文からノイズとなる語句を除去した後，名詞を抽出し文章の単語の重みベクトルを作成する．入力コーパスの辞書ベクトルは [“ 簡単 ”, ”協力 ”, ”バトル ”, ”仲間 ”, ”ギルド ”, ”勝利 ”] となり，各文の重みベクトルはそれぞれ，$v_1 = [1, 1, 1, 1, 0, 0]$ と $v_2 = [0, 0, 1, 1, 1, 1]$ になる．文章の類似度 $similarity$ は重みベクトルのコサイン類似度 $\cos(v_1, v_2)$ により算出されるので $similarity = 0.5$ となり，この値が閾値 θ 以上の場合はノード間が接続され，θ 未満の場合には接続されない．上記の処理を，入力コーパスの文章の全ペアに行うことで文章ネットワークを構築し，接続エッジ数の多いノードのうち上位 M 件を Description の候補として抽出する．

3.1.2 類義語による単語入替

前節での文章ネットワークによる重要文抽出では，1 広告主あたり 10 〜 20 程度しか文章を抽出できないため，生成できる Title の数が少なくなってしまう傾向にある．そこで本稿では，文章中の単語を関連語や類義語と入れ替えることにより，Description 候補数を拡張する手法を提案する．

ここで利用する関連語や類義語は，Wikipedia の文章コーパスから Mikolov らによって提案された word2vec[*1] [Mikolov 13a, Mikolov 13b] を用いて学習する．word2vec を用いる理由は，WordNet の利用に比べて，アプリ説明文特有の未知語や新語などに対応しやすいからである．また，文章コーパスを Wikipedia のみではなく，これまで配信された膨大な TD データを用いることで，よりスマートフォンアプリの TD に特化した類義語抽出も可能になるという拡張性も利点の一つである．

word2vec では，ある単語を Distributed 表現と呼ばれるベクトルへと変換し，言語モデルを構築する手法の一種である．Wikipedia の文章コーパスを word2vec に入力し，コーパス内の各単語 w の Distributed 表現 $C(w)$ を学習する．Distributed 表現 $C(w)$ を学習する際に，単語 w の前後に出現する単語群 $w_1, \ldots, w_t$ が重要で,t を文脈幅と呼ぶ．ゆえに，単語 w_1 と w_2 の類似度は，それぞれの Distributed 表現のコサイン類似度 $\cos(C(w_1), C(w_2))$ により求められる．この word2vec を用いて，Description 候補文中の各名詞に対して，それぞれコーパス内の全単語と類似度を算出し，上位 N 件の名詞を類義語として元の名詞と入れ替える処理を行う．これにより，Description 数を M，Description の平均名詞含有数を K_{noun} とすると，$M \cdot K_{noun} \cdot N$ 件の Description 候補を得ることが可能となる．

3.2 Title 部の生成

抽出された Description 候補文を入力として，Title の候補を生成する処理を以下に述べる．

3.2.1 制約下での部分文字列抽出

スマートフォンアプリの説明文は，口語表現が多く，文法や係受け構造を抽出しにくい．そこで本稿では，単純な部分文字列抽出をいくつかの制約のもとで行うことで，見出しとして使える文字列を生成する．具体的には，入力文章に対して，以下の手順で制約に基づいた部分文字列を抽出を行う．なお，形態素解析器は MeCab[*2] (IPADIC 辞書) を用いた．

*1 https://code.google.com/p/word2vec/
*2 http://mecab.sourceforge.net/

1. 入力文章を形態素解析器によって単語分割する.
2. 分割された単語のうち, 以下の制約を満たす単語を注目単語とする.
   - 第一品詞が名詞か接頭詞である
   - 第二品詞が接尾でも非自立でもない
3. 各注目単語に対して全体で 15 文字以内に収まるように後方の単語を順次接続していき, 以下の制約を満たす場合に, そこまでで得られた部分文字列を Title の候補として出力する.
   - 最後に接続した単語が第一品詞が名詞か動詞か終助詞である
   - 記号, 句読点, 名詞の助動詞語幹, 動詞の連用形・連用タ接続のいずれでもない
   - 全体で 10 文字以上の長さがある

3.2.2 体言止めへの変換

上記の手法では, 厳しい制約の下で部分文字列を抽出しているため, 1 件の文章から平均で 2 件程度しか Title を生成することができない. そこで Title の種類を増やすために, 部分文字列抽出で出力した Title 候補のうち, 動詞で終わるものを体言止めに変換する処理も合わせて行う. 以下の具体例に示されているように,「名詞 + 格助詞 + 動詞」という構成になっている Title 候補を「動詞 + 名詞」という形式への変換処理をすることで, 体言止め表現へと変換する.

(例) 悩み　を　打ち明けられる　→　打ち明けられる　悩み

4. 評価実験

4.1 実験概要

評価実験では,2 つのスマートフォンアプリの広告主 A と B に対して, それぞれのアプリストアの説明文を取得し, 提案手法を適用して Title と Description を生成した. 生成された Title リストと Description リストのそれぞれから, 入稿者が TD として利用できそうなものを選択して入稿した. 実際に選択された TD の例を示す.

広告主 A の TD は 2015 年 2 月に 10 日間, 広告主 B の TD は 2015 年 3 月に 4 日間,Google AdWords での配信を行った. 広告主 A では, 既存 TD2 件に対して提案手法で自動生成した TD を 2 件, 広告主 B では, 既存 TD2 件に対して提案手法で自動生成した TD を 1 件を同時期に無作為に配信した. ただし, 配信実験中に広告の入札価格は変化させていない.

この実験での各段階におけるパラメータは, 事前に予備実験を行い, 表 1 のように設定している.

パラメータ	変数	設定値
重要文抽出での上位件数	M	50
重要文抽出での閾値	θ	0.2
類義語抽出での上位件数	N	3
word2vec の文脈幅	t	7
word2vec の Distributed 次元数	d	300
word2vec の単語の最低頻度数	f	4

表 1: 実験時のパラメータ設定値

4.2 評価指標と比較対象

本稿での評価指標は,TD 作成工数の削減率, 生成された Title と Description の有効率, そして広告効果を表す CTR, の 3 点である.

4.2.1 広告効果

提案手法によって生成された TD がどれぐらいの広告効果を持つのかを配信で得られる $CTR_t = Click_t/Impression_t$ によって評価する. ただし, $Impression_t$ はある TD t の表示回数を, $Click_t$ は t のクリック回数をそれぞれ表す. CTR において, 新しく生成された TD と, 各アプリでもともと配信されていた既存の TD とを, 同時に配信し実績を比較することで提案手法の広告効果を検証する.

4.2.2 生成した候補リストの有効率

候補リストの有効率とは, 提案手法で出力される全ての TD 候補数に対する, 入稿者が TD に利用できそうだと感じた TD 数の割合を意味する.

$$\text{有効率}_l = \frac{l \text{ 内から実際に利用できる } TD \text{ 数}}{l \text{ 内の } TD \text{ 数}}$$

ただし, l は候補リストを表す. すなわち,有効率$_l$ は l 内にどれだけ広告テキストとして使える TD が含まれているかを表し, この指標で高い値をもつシステムは有用であると考えられる.

有効率の観点から提案手法の評価を行うため,Description の候補リストは以下の 4 パターン生成している.

- 重要文抽出で算出される単語ベクトルの要素に単語の出現有無を用いており,
  - 類義語による文章拡張を導入しないパターン D_1
  - 類義語による文章拡張を導入するパターン D_2
- 重要文抽出で算出される単語ベクトルの要素に単語の tfidf 値を用いており,
  - 類義語による文章拡張を導入しないパターン D_3
  - 類義語による文章拡張を導入するパターン D_4

Title の候補リストは, D_1〜D_4 に体言止め変換の有無の 2 パターンを追加して, 合計 8 パターン生成している.

- 体言止め変換を導入しないパターン D_nT_1
- 体言止め変換を導入するパターン D_nT_2

全てのパターンで Title と Description の候補リストを出力し, 検索連動型広告の運用経験を 2 年以上持つ運用業務従事者が TD として有効か否かを判定する. 各パターンごとに有効率を算出し, 各段階の手法がどの程度影響しているのかを比較する.

4.2.3 TD 作成工数の削減率

新規の広告主への TD を作成する工数として, 現状では,1 広告主あたり 8 人時（作業者 1 人として約 8 時間程度）の工数を必要とする. その内容は, 新規の広告主サービスの理解から始まり, 良い言い回しの列挙やこれまでの配信経験から得られた知見に沿う TD の考慮などである. 工数削減率は提案システムの導入時と未導入時の工数から以下の計算式で算出される. 工数の単位として“ 人時 ”を用いる.

$$\text{工数削減率} = \frac{\text{導入時の工数}}{\text{未導入時の工数}}$$

上記指標によって, 提案システム導入により工数の削減度合いを測る.

4.3 実験結果

以下に, 各指標における実験結果を示す.

4.3.1 CTR

広告主	TD	Impression	Click	**CTR**
A	既存 TD1	174772	202	0.0012
A	既存 TD2	239523	315	0.0013
A	提案 TD1	229084	278	0.0012
A	提案 TD2	274967	349	0.0013
B	既存 TD3	6242	407	0.065
B	既存 TD4	3713	43	0.012
B	提案 TD3	3456	107	0.031

表 2: 各 TD の配信実績

配信の結果, 広告主 A と B に対する, 既存 TD と提案 TD の実績は表 2 のようになった. 広告主 A においては, 既存 TD と提案 TD の間で CTR にほぼ差はなく, 広告効果に大きな変化がないことが確認できた. また, 広告主 B に対しては, 自動生成した提案 TD の CTR が 2 つの既存 TD の平均に近い値となっている. 広告主 B の実績については, 配信期間が短いため速報値として傾向を見るために扱うが, その傾向としても大きく CTR が劣っている状態ではないことが確認できた.

4.3.2 TD リストの有効率

広告主 A に対して自動生成システムが出力した Title と Description のリストの有効率は, それぞれ表 3 と表 4 のようになった.

D_mT_n	生成件数	有効件数	**有効率**
D_1T_1	59	15	0.254
D_1T_2	91	17	0.187
D_2T_1	486	20	0.041
D_2T_2	689	21	0.030
D_3T_1	38	7	0.184
D_3T_2	60	7	0.117
D_4T_1	336	9	0.027
D_4T_2	480	9	0.019

表 3: 各手法の組み合わせごとの Title リストの有効率

D_n	生成件数	有効件数	**有効率**
D_1	43	5	0.116
D_2	562	10	0.017
D_3	22	13	0.590
D_4	247	32	0.130

表 4: 各手法の組み合わせごとの Description リストの有効率

まず, 単語拡張を適用すると生成件数を約 8 〜 10 倍増やすことができている. しかし,Title,Description ともに, 類義語による文章拡張を行うと有効率が著しく下がり, 約 0.1 〜 0.15 倍になるということが分かった. また, 体言止め変換の適用では生成件数を約 1.5 〜 2 倍増やすことができているが, それに対して有効件数が増えてないため, 有効率が約 0.5 〜 0.75 倍になっている.

4.3.3 TD 作成の工数

提案システムを用いた場合, アプリストアの説明文を与えさえすれば, 瞬時に TD リストが出力される. 出力された TD リストを人手でチェックする時間が必要となる. 本実験では, どのパターンの TD においても, 1 人時程度の工数でチェックを完了できることが計測できた. ゆえに, 提案手法による工数削減率は $\frac{1}{8}$ であった.

5. おわりに

本稿では, スマートフォンアプリのストアの説明文から広告テキストを自動生成する手法を提案した. 提案手法では, 文章ネットワークを用いた重要文抽出や類義語抽出による単語拡張によって Description の候補を抽出し, 制約下での部分文字列抽出や体言止め変換によって Title の候補を生成した.

提案手法で生成した TD の CTR は既存の TD に比べて大きな変化がなく, 自動生成した TD によって広告効果が著しく下がることはないといえ, また,TD の作成工数を $\frac{1}{8}$ に削減できることが分かっているため, 提案手法では広告効果を保ちつつ, 実作業時間を削減することが可能である.

しかし,TD リストの有効率に関しては, 全体的に非常に低く, 特に TD の種類を増やすために提案した, 単語拡張や体言止め変換を導入した場合に著しく低下しており, 有効な TD 数が増える割合よりも人手でチェックする TD 数が増える割合の方が高くなってしまっていることは改善すべき点である.

今後は有効率を上げるために, 類義語学習におけるコーパスにアプリ説明文データやこれまで配信した TD データを加えて類義語学習の精度を高めることや, 体言止め表現への変換により強い制約を加えることなどに取り組んでいきたい.

参考文献

[Mikolov 13a] Mikolov, T., Chen, K., Corrado, G., and Dean, J.: Efficient Estimation of Word Representations in Vector Space, *CoRR*, Vol. abs/1301.3781, (2013)

[Mikolov 13b] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J.: Distributed representations of words and phrases and their compositionality, in *Advances in Neural Information Processing Systems*, pp. 3111–3119 (2013)

[Salton 96] Salton, G., Singhal, A., Buckley, C., and Mitra, M.: Automatic Text Decomposition Using Text Segments and Text Themes, in *Proceedings of the the Seventh ACM Conference on Hypertext*, HYPERTEXT '96, pp. 53–65, ACM (1996)

[幾島 08] 幾島 克洋, 藤田 篤, 佐藤 理史, 横川 睦, 岩本 宜式, 片岡 亮：HTML 文書からのリスティング広告の自動生成, 言語処理学会第 14 回年次大会発表論文集, pp. pp.504–507 (2008)

[藤田 11] 藤田 篤, 幾島 克洋, 佐藤 理史：検索連動型広告の自動生成と集客効果の測定-飲食店ドメインを例題に, 情報処理学会論文誌, Vol. 52, No. 6, pp. 2031–2044 (2011)