Vol.2018-NL-236 No.5
2018/7/9

# Distant Supervisionによる質問要約

石垣 達也[1,a)]　町田 和哉[1,b)]　小林 隼人[2,3,c)]　高村 大也[4,5,d)]　奥村 学[4,e)]

**概要**：学会での質疑応答や電子メールでのやりとりにおいて質問は広く用いられている．このような質問には，核となる質問の他に補足的な情報も付与され，ときに冗長になることがある．本研究では冗長な質問をその内容を端的に表現する 1 文の質問に要約する課題を扱う．既存研究では，コミュニティ質問応答サイトに投稿される質問の本文とタイトルの対を長文質問とその要約とみなし，教師あり要約モデルを構築する手法が提案されている．しかし，例えば日本語の Yahoo!知恵袋などにおいてはタイトルが存在せず，先行研究のように擬似的な教師データを獲得することは困難である．そこで、本研究では，Distant Supervision に基づく抽出型質問要約器を提案する．評価実験において，提案要約器が教師あり及び教師なし学習に基づくベースライン手法を上回る性能を示した．

# Distant Supervision for Question Summarization

Ishigaki Tatsuya[1,a)]　Machida Kazuya[1,b)]　Kobayashi Hayato[2,3,c)]　Takamura Hiroya[4,5,d)]
Okumura Manabu[4,e)]

## 1. はじめに

学会での質疑応答や電子メールでのやりとりなどにおいて，我々はしばしば質問を投げかける．このような場面での質問には，核となる質問の他に補足的な情報も付与され，ときに冗長になることがある．補足的な情報は質問内容を完全に理解するには必要であるが，要旨を素早く把握したいといった状況においては，かえって理解の妨げになる．そこで，本研究では冗長な質問をより短く理解しやすい単一文質問に要約する課題に取り組む．

文書要約は自然言語処理分野において長く研究されている課題の一つである．文書要約に用いられる手法は抽出型と生成型に大別される．抽出型手法は入力文書から文や単語などの単位を選択し，つなぎ合わせることで要約を出力する [5], [10], [13], [17], [18], [24]．一方，生成型手法は入力文書の表現を言い換えするなど，入力に含まれない表現も用いながら要約を生成する [4], [26]．生成型手法は圧縮率の観点に置いては有利ではあるが，抽出型手法に比べ文法誤りを含む文を出力しやすい傾向にある．質問応答サイトのタイトルなど文法誤りの許されない実応用においては，誤りの少ない抽出型手法による要約の利点は大きい．よって，本研究では質問要約課題を抽出型要約として解く．

既存の要約研究が対象とするテキストは，ニュース記事 [23]，商品レビュー [27]，脚本 [9]，メールスレッド [3]，会話ログ [21]，議事録 [22] など多岐に渡る．しかしながら，質問を明示的に対象とした要約研究は少ない．田村ら [25] は，質問応答システムの性能を向上させる目的で，複数文で構成される質問から核文を同定する分類器を学習した．この分類器は 2,000 文書に対し重要文を人手でアノテーションしたデータから学習されており，データ作成のコストが大きい．Ishigaki ら [12] は質問応答サイトである Yahoo! Answers[*1] に投稿される質問とそのタイトルの対を長文質問とその要約の対とみなし，英語で記述された質問を対象とする抽出型および生成型の要約モデルを学習した．

1 東京工業大学大学院
2 ヤフー株式会社
3 理化学研究所 AIP センター
4 東京工業大学科学技術創成研究院
5 産業技術総合研究所
a) ishigaki@lr.pi.titech.ac.jp
b) machida@lr.pi.titech.ac.jp
c) hakobaya@yahoo-corp.jp
d) takamura@pi.titech.ac.jp
e) oku@lr.pi.titech.ac.jp

[*1] http://answers.yahoo.com

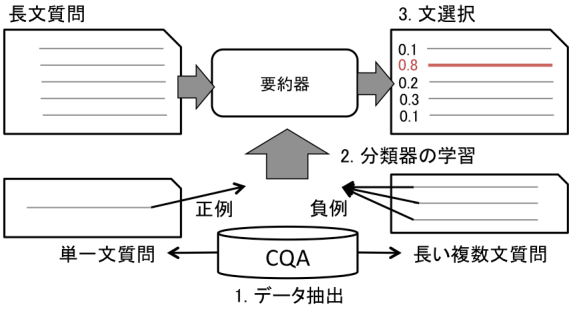


**図 1** Distant Supervision による質問要約の概要.

質問応答サイトのタイトルは新聞記事などに付与されるタイトルとは異なり，専門的な訓練を受けていないユーザが自由に付与したものであり，ノイズが多い．また，Yahoo!知恵袋や Quora などの質問応答サイトにはそもそも質問投稿時にタイトルを入力しない．したがって，例えば日本語の質問を対象とした教師あり要約モデルを学習するために必要な質問とタイトルの対を獲得するのは困難である．

教師データの獲得が難しい状況において，Distant Supervision による分類器構築がいくつかの自然言語処理課題向けに提案されている．Mints ら [19] によって，はじめて提案された Distant Supervision は，ヒューリスティックや規則，外部データなどを用いて，コストをかけずに大量のラベル付きデータを生成し学習する手法である．Mints ら [19] はテキスト中に出現する 2 つのエンティティの関係を抽出する分類器を Distant Supervision により学習した．この課題では，例えば，“Steve Jobs is the CEO of Apple.” という文を入力として受け取り，“Founder(Steve Jobs, Apple)” といったエンティティと関係を出力する．このような分類器を学習するには Steve Jobs や Apple がエンティティであるいった情報や，これらのエンティティが Founder の関係であるといった情報がアノテーションされたデータを作成する必要があるが，人手作成にはコストがかかる．そこで，Mints らは知識ベースの一つである Freebase[2] に格納された情報を手がかりに，テキストに対し自動でエンティティや関係ラベルを付与し，擬似的な教師データを作成した．このほかにも，絵文字を手がかりにツイートに対し極性ラベルを付与する手法 [8]，Wikipedia のトピックラベルを手がかりにブログ記事にカテゴリラベルを自動付与する手法 [11] など，多様なヒューリスティックや規則，外部データが様々な課題向けに提案されている．

要約課題においては，人間が付与した新聞記事の要約を手がかりにニューラルネットを用いた抽出型要約向けの教師データを自動作成する手法 [5], [20] が提案されており，Distant Supervision の一種であると考えられる．英語など，一部の言語資源が豊富な言語に対してはこのようなデータ作成手法は有益であるものの，日本語など質問-要約対の獲得が難しい言語への適用は難しい．

| (a) 単一文質問 |
|---|
| ヤフーオークションのプレミアムアカウントはどうやったら取れますか？ |
| (b) 複数文質問 |
| オークション初心者です．ヤフオクについてあまり詳しくありません．自分の商品を何人が見たか気になっています．アクセス総数には自分も含まれますか？同じ人がアクセスしたらどうなりますか？ |

**表 1** 単一文質問と複数文質問の例

そこで，本研究では質問-要約対を必要としない Distant Supervision による質問要約器を提案する．提案要約器は Yahoo!知恵袋に投稿される単一文質問と要約に含めるべき文が似た特徴を持つことに着目し，質問-要約対なしで要約器を学習する．クラウドソーシングを用いて人手作成した評価セットによる比較実験において，Distant Supervision による提案要約器が，教師ありおよび教師なし学習によるベースライン手法よりも，要約に含めるべき文を正しく同定できたをこと報告する．

## 2. 質問要約のための Distant Supervision

質問要約課題において，出力は 1) 質問であること，2) 他の情報を利用せず出力のみで文意が理解できることの 2 つの条件を満たす必要がある．例えば，表 1 に示す “ヤフーオークションのプレミアムアカウントはどうやったら取れますか” という質問には，“ヤフーオークション” や “プレミアムアカウント” といった，質問内容を理解するために必要な情報があらかじめ埋め込まれており，この文の情報のみで文意を理解することができる．一方で，質問応答サイトに投稿される長い複数文質問に含まれる文それぞれは，理解するために他の文の情報を必要としたり，そもそも質問文でさえない場合が多い．例えば，表 1 の複数文質問の例において，“アクセス総数には自分も含まれますか？”という文のみを提示された場合，“オークション”などの重要な情報が欠如しこの文のみでは質問が理解できない．このような文は質問に含めるべき文とは異なる特徴を持つ．Yahoo!知恵袋データセットを用いた分析では，無作為に抽出した 100 の単一文質問のうち 98%はその文以外の情報を利用せず内容が理解できる質問であった．一方で，10 文以上で構成される長い複数文質問に含まれる文を人手で分析すると，わずか 7%がその文以外の情報を利用せず内容が理解できる質問であった．よって，質問応答サイトに投稿される単一文質問のほとんどは，その文以外の情報を利用せず内容が理解できる質問であり，質問要約課題の出力と似た特徴を持っていることがわかった．このような分析により我々は，要約に含めるべき文は，単一文質問か長い複数文質問中の文であるかを分類する分類器を用いて同定できると考えた．

## 3. 提案手法

以上の仮説に基づき，図 1 に示す Distant Supervision による質問要約の枠組みを提案する．この枠組みでは，まず質問応答サイトの投稿から単一文質問及び，10 文以上で構成される質問を抽出する (手順 1)．次に，単一文質問を正例，10 文以上で構成される質問中の文を負例として，分類器を学習する (手順 2)．最後に，この分類器の出力する確信度を文の重要度とみなし，この重要度スコアを基に抽出文を決定する (手順 3)．

### 3.1 データ抽出 (手順 1)

データ抽出の元となるコミュニティ質問応答には Yahoo! 知恵袋データセット第 1 版[*2]を用いた．このデータには，2004 年 4 月から 2005 年 10 月に収集したおよそ 300 万件の質問投稿が含まれる．このデータから，単一文質問を 8,000 文，10 文以上から構成される複数文質問中の文を 1,700,000 文，無作為に抽出した．抽出文のうち，単一文質問を正例，10 文以上から構成される複数文質問中の文を負例としてロジスティック回帰モデルの学習に用いる．

### 3.2 分類器の学習 (手順 2)

文の重要度を出力する分類器について説明する．分類器にはロジスティック回帰モデルを用い，単一文質問を正例，10 文以上から構成される複数文質問中の文を負例とし学習する．ロジスティック回帰モデルが出力する確率値 $p(s) = \sigma(\boldsymbol{\theta}^\top \boldsymbol{f}(s))$ を入力文書中の文 $s$ に対する重要度スコアと考え抽出文を決定する．ここで $\boldsymbol{\theta}$ は学習パラメータ，$\boldsymbol{f}(s)$ は文 $s$ の素性ベクトル，$\sigma$ はシグモイド関数で $\sigma(x) = 1/(1+e^{-x})$ で定義される．

### 3.3 文選択 (手順 3)

次に分類器の出力確率を用いた文選択手法について述べる．入力質問には複数の質問が含まれる場合がある．そのような場合には，どちらのフォーカスがより重要であるか判定するのは難しい．本研究では，入力質問が複数のフォーカスを持つ場合には，より前に出現するフォーカスを正解として扱う．

そのため，分類器の出力確率が閾値 $\tau_{\text{dist}}$ 以上の文が複数存在する場合には，入力に複数のフォーカスが存在するとみなし，より前に出現する文を出力する．それ以外の場合には確率が最大となる文を出力する．

## 4. 実験

本節では実験に用いたデータ，提案モデルおよびベースライン，比較実験について述べる．

[*2] http://www.nii.ac.jp/dsc/idr/en/yahoo/yahoo.html

| 質問の文数 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|
| 事例数 | 461 | 349 | 199 | 91 | 42 | 38 |

**表 2** 評価に用いた事例数とその事例内の文数.

### 4.1 評価セット

評価セットの作成においては，2 文から 7 文で構成される複数文質問を Yahoo!知恵袋データセットから 2,000 件抽出し人手により正解抽出文をアノテーションした．なお，分類器の学習には単一文質問および 10 文以上から構成される質問を用いているため，学習データと評価データでの重複はない．クラウドソーシングサービスである Lancers[*3]を用いて，要約に含めるべき文を作業者にアノテーションしてもらった．5 人の作業者に入力質問のうち少なくとも 1 文を選択してもらい，より短い質問の生成を行った．作業者には以下の手順での要約の作成を依頼した．

(1) 質問全体を読み質問のフォーカスをよく考える．
(2) 核となる質問文を選ぶ．
(3) 内容が理解できる要約になるまで追加で補足文を選ぶ．

4 人以上の作業者が要約に含めると判定した文を正解抽出文とみなした．アノテーションされたデータのうち，1 文のみが正解抽出文となった 1,180 文書を評価セットとした．1,180 文書中の文数の分布を表 2 に示す．

### 4.2 性能評価

実装には LIBLINEAR [7] の L1 正規化項付きのロジスティック回帰を用いた．素性ベクトルの構築には単語および品詞タグの unigram，bigram，trigram を用いた．単語分割と品詞タグ付けには MeCab [15] を用いた．MeCab のシステム辞書には IPADIC [1] を用いた．コストパラメーター $C$ は訓練データ内の 5 分割交差検定において，以下のように定義される正解率が最大になる値に設定した:

$$\text{正解率} = \frac{\text{抽出文を正しく同定できた文書数}}{\text{評価に用いた文書数}}$$

### 4.3 ベースライン

Distant Supervision による提案要約器を，教師あり学習，教師なし学習に基づくベースライン手法と比較した．教師あり学習に基づくベースライン手法は，評価セットを 5 分割交差検定し正解率を計算した．具体的には，評価セットを 3:1:1 に分割し，それぞれをモデルの学習，パラメータチューニング，正解率の計算に用いた．

教師あり学習によるベースラインとしては，人手アノテーションされたデータから学習したロジスティック回帰モデルによる分類手法を用いる．`Super` はロジスティック回帰モデルによる教師あり学習手法である．モデルの学習と評価には評価セットに付与され人手によるアノテーショ

[*3] http://lancers.co.jp

| 手法/文数 | 2 | 3 | 4 | 5 | 6 | 7 | All |
|---|---|---|---|---|---|---|---|
| DistInit | **.92** | **.90** | **.82** | **.80** | **.74** | **.71** | **.81** |
| Dist | .84 | .85 | .75 | .70 | .65 | .59 | .73 |
| Lead-Q | .89 | .83 | .77 | .63 | .58 | .57 | .72 |
| Lead | .88 | .73 | .58 | .58 | .53 | .44 | .62 |
| LexRank | .88 | .70 | .57 | .47 | .42 | .32 | .56 |
| SimEmb | .77 | .60 | .50 | .31 | .31 | .21 | .45 |
| TfIdf | .68 | .45 | .36 | .28 | .22 | .19 | .36 |
| SuperInit | .90 | .84 | .73 | .75 | .53 | .51 | .71 |
| Super | .84 | .78 | .75 | .75 | .50 | .43 | .67 |

**表 3** 各手法の正解率．最上段は入力文書中の文数．

ンを用いた．素性ベクトルの構築には，Distant Supervision に基づく提案モデルと同様の素性を用いた．提案モデルと同じ条件で比較を行うために，このロジスティック回帰モデルが出力する確率値が閾値 $\tau_{\text{super}}$ 以上の文が複数ある場合にはより先頭を選択する SuperInit とも比較する．閾値は 5 分割交差検定においてパラメータチューニング用の分割データでの正解率が最大となる値 $\tau_{\text{super}} = 0.5$ を採用する．

教師なし学習に基づく手法としては，2 つの規則に基づく手法，3 つの統計的手法と比較する．Lead は先頭文を選ぶベースラインであり，一般的な要約課題では強いベースラインとして知られている．Lead-Q は先頭質問文を出力するベースラインであり，質問要約課題においては Lead よりも強いベースラインである [12]．質問文の同定には Tamura ら [25] の例示している規則を用いた．LexRank [6] はよく知られたグラフに基づくベースラインである．TfIdf は単語あたりの TF-IDF の値が最大となる文を選択する．IDF は Yahoo!知恵袋の質問投稿全体から計算した．SimEmb は入力文書と抽出文の意味的な距離が最小となる文を選択するベースラインである．意味空間における距離は，教師なし学習に基づく既存要約研究ではよく用いられる [13], [14]．距離の計算尺度には TF-IDF によるコサイン類似度よりもよい性能を示している Word Movers' Distance (WMD) [16] を用いた．

## 5. 結果と考察

各手法の正解率を表 3 に示す．DistInit は Distant Supervision に基づく提案モデルである．パラメーターチューニング用の分割データでの性能が最大になるよう閾値を $\tau_{\text{dist}} = 0.7$ に設定した．Dist は Distant Supervision に基づく分類器の出力確率が最大の文をナイーブに選択するモデルである．

結果として，DistInit が比較手法の中でもっとも良い性能を示した．DistInit (.81) と Lead-Q (.72) との差は符号検定を用い，統計的有意 ($p < 0.05$) であることを確認した．Dist は強いベースラインとして知られる Lead-Q よりもわずかに良い性能を示した．Dist は入力文書が複数のフォーカスを持つ場合に出力に失敗しており，より先頭のフォーカスを選ぶ DistInit によって性能が上がった．

| スコア | 文 |
|---|---|
| .46 .08 .22 | 質問の意味を勘違いしているよう。<br>ですので 再掲載します |
| .95 .37 .10 | **何故日本ではディーゼル車を<br>推進しないのか?** |
| .90 .09 .11 | 国産メーカでも優秀なディーゼル<br>エンジン車、 海外で販売してるじゃないか! |
| .82 .02 .12 | ハイブリット化ってバッテリー化推進してる<br>けどバッテリー作る時も使い終わった<br>廃バッテリー処理するのに、どんなに<br>環境破壊してのかと思っているのですが。 |
| .91 .58 .11 | 実際はどちらが環境によいの? |
| .35 .60 .28 | バッテリーを生産 処分する過程に起きる<br>環境問題も含めて聞きたいです。 |

**表 4** 実際に付与された重要度スコアの例. 数値は左から順にそれぞれ，Dist ，Super ，SimEmb によるスコアを示す. **太字は提案手法 DistInit の選択した文.**

実際の出力例を表 4 に示す．Distant Supervision に基づく分類器は 2 文目や 5 文目の質問文に対し高いスコアを付与した．提案モデルでは "!" などの記号にも高い重みが付いており，3 文目のように質問文以外の文に高いスコアが与えられることもある．DistInit は複数のフォーカスを含む文に対し Dist よりも正解文を選ぶことが多く，性能の向上につながった．教師なし学習に基づくモデルや教師あり学習に基づくモデルの出力は必ずしも質問文ではないことが，高い性能を示さない要因となったと考えられる．

## 6. まとめ

本稿では Distant Supervision による質問要約手法を提案した．提案モデルは人手によるアノテーションなしで，教師ありおよび教師なし学習に基づくベースライン手法よりも良い性能を示した．複数のフォーカスを持つ入力に対し相対的な重要度を計算するモデルへの拡張を検討している．また，質問要約に限らずより一般的な新聞記事などの文書要約課題，複数文書要約課題において Distant Supervision を適用することが可能であるか検証したいと考えている．

### 参考文献

[1] Asahara, M. and Matsumoto, Y.: ipadic version 2.7. 0 User's Manual, *Computational Linguistics Laboratory. Graduate School of Information Science. Nara Institute of Science and Technology* (2003).
[2] Bollacker, K., Evans, C., Paritosh, P., Sturge, T. and Taylor, J.: Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge, *Proceedings of SIGMOD2008*, pp. 1247–1250 (2008).
[3] Carenini, G., Ng, R. T. and Zhou, X.: Summarizing Emails with Conversational Cohesion and Subjectivity, *Proceedings of ACL-08: HLT*, pp. 353–361 (2008).
[4] Chen, Q., Zhu, X., Ling, Z., Wei, S. and Jiang, H.: Distraction-based neural networks for modeling documents, *Proceedings of the IJCAI2016*, pp. 2754–2760 (2016).
[5] Cheng, J. and Lapata, M.: Neural Summarization by Extracting

Sentences and Words, *Proceedings of ACL2016*, Berlin, Germany, pp. 484–494 (2016).
[6] Erkan, G. and Radev, D. R.: Lexrank: Graph-based lexical centrality as salience in text summarization, *journal of artificial intelligence research*, Vol. 22, pp. 457–479 (2004).
[7] Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R. and Lin, C.-J.: LIBLINEAR: A Library for Large Linear Classification, *Journal of Machine Learning Research* (2008).
[8] Go, A., Bhayani, R. and Huang, L.: Twitter Sentiment Classification using Distant Supervision, *Stanford Technical Report* (2009).
[9] Gorinski, P. J. and Lapata, M.: Movie Script Summarization as Graph-based Scene Extraction, *Proceedings of NAACL2015*, pp. 1066–1076 (2015).
[10] Hirao, T., Isozaki, H., Maeda, E. and Matsumoto, Y.: Extracting Important Sentences with Support Vector Machines, *Proceedings of COLING2002*, pp. 1–7 (2002).
[11] Husby, S. D. and Barbosa, D.: Topic classification of blog posts using distant supervision, *Proceedings of the Workshop on Semantic Analysis in Social Media*, pp. 28–36 (2012).
[12] Ishigaki, T., Takamura, H. and Okumura, M.: Summarizing Lengthy Questions, *Proceedings of IJCNLP2017*, Vol. 1, pp. 792–800 (2017).
[13] Kågebäck, M., Mogren, O., Tahmasebi, N. and Dubhashi, D.: Extractive summarization using continuous vector space models, *Proceedings of CVSC2014*, pp. 31–39 (2014).
[14] Kobayashi, H., Noguchi, M. and Yatsuka, T.: Summarization based on Embedding Distributions, *Proceedings of EMNLP2015*, pp. 1984–1989 (2015).
[15] Kudo, T.: Mecab: Yet another part-of-speech and morphological analyzer, *http://mecab. sourceforge. jp* (2006).
[16] Kusner, M., Sun, Y., Kolkin, N. and Weinberger, K.: From word embeddings to document distances, *Proceedings of ICML2015*, pp. 957–966 (2015).
[17] Luhn, H. P.: The Automatic Creation of Literature Abstracts, *IBM J. Res. Dev.*, Vol. 2, No. 2, pp. 159–165 (1958).
[18] Mihalcea, R. and Tarau, P.: TextRank: Bridging Order into Texts, *Proceedings of EMNLP2004*, pp. 404–411 (2004).
[19] Mintz, M., Bills, S., Snow, R. and Jurafsky, D.: Distant supervision for relation extraction without labeled data, *Proceedings of ACL-IJCNLP2009*, pp. 1003–1011 (2009).
[20] Nallapati, R., Zhai, F. and Zhou, B.: SummaRuNNer: A Recurrent Neural Network Based Sequence Model for Extractive Summarization of Documents., *Proceedings of AAAI2017*, pp. 3075–3081 (2017).
[21] Oya, T. and Carenini, G.: Extractive Summarization and Dialogue Act Modeling on Email Threads: An Integrated Probalistic Approach, *Proceedings of SIGDIAL2014*, pp. 133–140 (2014).
[22] Oya, T., Mehdad, Y., Carenini, G. and Ng, R.: A Template-based Abstractive Meeting Summarization: Leveraging Summary and Source Text Relationships, *Proceedings of INLG2014*, pp. 45–53 (2014).
[23] Rush, A. M., Chopra, S. and Weston, J.: A Neural Attention Model for Abstractive Sentence Summarization, *Proceedings of EMNLP2015*, pp. 379–389 (2015).
[24] Takamura, H. and Okumura, M.: Text summarization model based on maximum coverage problem and its variant, *Proceedings of EACL2009*, pp. 781–789 (2009).
[25] Tamura, A., Takamura, H. and Okumura, M.: Classification of Multiple-Sentence Questions, *Proceedings of IJCNLP2005*, pp. 426–437 (2005).
[26] Tan, J., Wan, X. and Xiao, J.: Abstractive document summarization with a graph-based attentional neural model, *Proceedings of ACL2017*, Vol. 1, pp. 1171–1181 (2017).
[27] Yu, N., Huang, M., Shi, Y. and zhu, x.: Product Review Summarization by Exploiting Phrase Properties, *Proceedings of COLING 2016*, pp. 1113–1124 (2016).