社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

第二種研究会資料
IEICE SIG Notes
WI2-2009-20(2009.3)

Grammar-gram と GrammarVerb-gram を用いたドメイン非依存型 Why テキストセグメント判定と回答抽出

田中 克幸† 滝口 哲也‡ 有木 康雄‡

†‡神戸大学工学部〒657-8501 神戸市灘区六甲台町1－1 神戸大学 自然科学研究棟3号館
E-mail: †katsutanaka@me.cs.scitec.kobe-u.ac.jp, ‡{takigu, ariki}@kobe-u.ac.jp

あらまし 本研究では，テキストセグメントの文法情報に着目して，ドメイン依存性の少ない学習器の構築により，Why テキストセグメント判定と回答抽出の自動化手法を提案し，1度の学習とラベル付けで様々なドメインにおける Why 型質問応答が可能となるグローバル学習器の構築を提案する．これにより従来の Why 型質問応答の問題点であった，ルール作成に手間が掛かる，ドメイン依存性，学習時間が長いといった問題が改善された．

キーワード QA, IR, Non-factoid

Domain Independent Why Text Segment Classification and Answer Extraction by Grammar-gram and GrammarVerb-gram

Katsuyuki TANAKA† Tetsuya TAKIGUCHI‡ and Yasuo Ariki‡

†‡Faculty of Engineering, First University 657-8501 Kobe University, Nada, Kobe 657-8501, Japan
E-mail: †katsutanaka@me.cs.scitec.kobe-u.ac.jp, ‡{takigu, ariki}@kobe-u.ac.jp

Abstract In this paper, we focus on constructing a global classifier to perform automated Why Text Segments Classification and Why Answers Extraction by proposing Grammar-gram and GrammarVerb-gram. The experimental results showed promising results on global ability, flexibility, effectiveness, efficiency of the classifier on Why Text Segments Classification and Why Answer Extraction. Hence it solved some of the problems Why-QA was facing.

Keyword QA, IR, Non-factoid

1. はじめに

近年，質問応答技術は，ユーザー負担を軽減する検索方法としてその有効性が注目され，TREC QATrack3[1] や NTCIR4 の QAC タスク[2]など，質問応答システムを評価するための評価プロジェクトも実施されて，多くの研究が行われている[5], [7], [8], [11].

しかし、従来の質問応答システムは，事実を対象とした質問に回答する研究，つまり factoid 型質問応答システムが主流である．“～はなぜ？”のように原因を求める Why 型質問応答(Why-QA)を探究する non-factoid 型 QA の研究例は少ない．

Why-QA に関する従来研究としては，原因・理由を特徴付ける幾つかのキーワードをもとに，回答を抽出するためのルール辞書を作成しておき，Why 型質問(Why-Q)に対して，原因・理由を表している文から回答箇所(Why-A)の抽出を行なうルールベース手法が一般的である[4], [6], [9], [12], [13]. ルールベース手法は，ドメインに依存せずに Why-QA が行なえるが，ルール作成の手間と，ルールの網羅性不足による Why-A の回答判定精度に問題がある．

統計的学習手法などを用いて，non-factoid 型や Why 型質問の解析を行い，FAQ ペアなどにより，質問に対して既に用意されている回答を提供するといった方法で，Why 型の質問に対応する研究も行なわれている[1], [10], [14], [15]. しかし，これらの手法では，任意のドキュメント集合から Why-QA の答を含んでいるドキュメントを特定することや，特定されたドキュメントの集合から Why 型のテキストセグメントを判定し，質問の回答を特定することについて言及されていないものが多い．また，non-factoid 型の質問に対する回答が，factoid 型質問程度の回答精度で行えないこと，更に，研究が Why-QA に特化されていない等といった問題点が挙げられる[1], [10].

機械学習を用いて Why 型のテキストセグメントを判定し，回答を抽出する方法も提案されているが[2]，この手法は，ドメインごとに学習を行わなければならず学習に時間がかかり，抽出された情報にも，対象ドメイン固有の表現が含まれ，作成された学習器はドメイン依存性が強いといった問題がある．

近年のインターネット情報の爆発的な拡大に伴い，

[1] http://trec.nist.gov/
[2] http://research.nii.ac.jp/ntcir/

factoid 型 QA のみならず，Why-QA の技術も重要である．特に，Why-QA に特化された技術と，インターネット上のドメイン制限のないダイナミックな世界において，これらの技術の確立が必要である．

そこで，本研究では，従来の Why-QA の問題点である，ルール作成の手間，ルールの網羅性，学習速度，ドメイン依存性といった問題を解決し，インターネット上にあるテキスト文書に対して Why 型のテキストセグメントを判定する **Why テキストセグメント判定(Why-TSC)**と，判定されたテキストセグメント(TS)から質問に対する原因・理由となる回答抽出(Why-AE)の自動化を目的としている．特に，本研究では，どの TS にも頻繁に現れて抽象度が高く，Why の特徴を多く含む文法情報に着目し，機能語を主として TS を表現した Grammar-gram (G-gram)と，G-gram に動詞も含めて TS を表現した GrammarVerb-gram (GV-gram)手法を提案し，ドメインに依存せず，これらの判定が行なえる手法を明らかにする．本研究では，Why-QA を分類問題と捉え，G/GV-gram 手法を用いて機械学習を行なうことで，Why-TSC と Why-AE を行なっている．機械学習を用いることで，手動によるルール辞書作成の手間を省き，ルールの網羅性を向上できる．また，機械学習において文法的知識を用いることにより，ドメインに依存しない，“グローバルな学習器”が高速に作成できるという利点がある．本手法の評価は，従来研究である 1).ルールベース手法[4]と，2).形態素 Uni-gram による機械学習を用いた手法[2]の 2 手法をベースラインとして，提案手法と比較する．

本論文の構成は以下の通りである．2 章では，関連研究を述べ，3 章では Why-TS と Why-AE について述べ，4 章では提案手法の Grammar-gram，GrammarVerb-gram を述べる．5 章で提案手法の評価実験結果と考察を示す．

2. 関連研究

渋沢ら[4]は，Web 文書を対象としたオープンドメインで，ルールベース手法を用いて Why-Q に対する Why-A の抽出を行った．質問文中のキーワードを全て含んでいる Web 文書中の 1 文を“事実文”と定義し，事実文を含む TS に対して，事実文中，又は事実文前後の文の何処に事実に対する原因・理由が現れるかを判別可能にした．この手法では，予め原因・理由を表す語や指示語のルール辞書を用意し，これらの特徴語のパターン化を行い，回答の有無を判定し抽出を可能にした．しかし，この手法では，ルール辞書の構築に手間が掛かかることや，Why-QA としての特徴語を発見するためのルール不足によって，精度が低下するといった問題点が挙げられる．また，総合的に TS の Why 判定と回答抽出に対する適合率と再現率が低く，Why-QA としての精度が良いとは言いがたい．

これらのルールベースの問題点を改善するために，Why-QA ルールの自動抽出を行う方法も提案されている[19]．Higashinaka らは，専門家による手動でラベリングされた EDR 辞書を用いて“原因”を表すタグの付いた文を抽出し，抽出された文を Cabocha[6]等を使い，文の構造化と抽象化を行なってルールを作成した．作成されたルールをもとに，ランキングを学習手法を用い，文または段落のランキングによって Why-QA を行なっている．EDR 辞書はラベルの信用性は高いが，これらの辞書の更新は困難なので EDR 辞書の知識に制限され，新規ルールの再学習が困難であり，この手法では，回答抽出は言及されていない．

回答抽出の研究は，Verberne によって集中的に行なわれている[14]，[15]，[21]．Verberne は，既知の QA ペアを使って構造化を行い，Why-Q に対する回答抽出方法を提案している．しかし，構造化が複雑かつ大量のコーパス必要とするため手間が掛かる．

田中ら[2]は，ルールベースの網羅性の問題を解決する手法の 1 つとして，TS データに対して形態素 Uni/Bi-gram を素性として機械学習を行い，Why-TSC と Why-AE の自動化について研究を行っている．しかし，この方法は，ドメインごとに学習を行わなければならず，再学習，再ラベリングを行う手間がかかり，ドメイン依存性の強い方法である．また，名詞は日々変り，増加するため，学習速度も学習量の増加に伴い遅くなる．この手法は，クローズドメインにおける Why-TSC には効果的であるが，ドメイン内に頻出する名詞も学習するため，ドメイン間の情報のミスマッチがおき，オープンドメインでの判定には適しておらず，ドメイン制約に関して問題がある．

3. Why-TSC と Why-AE アルゴリズム

Why-QA タスクのために，インターネットに存在するフリーテキストの中から，Why-Q に対する Why-A を抽出する．このための重要なポイントは，次の 3 点である[2]．1)ドキュメント中の **“Why テキストセグメント判定”**．2)判定されたセグメントに対する **“事実・結果（質問）セグメント特定”**．3)見つけたセグメント内の **“理由・原因となる回答抽出”**．

Why テキストセグメント判定(Why-TSC)とは，任意のドキュメント D をセグメント化した TS 集合 $\{TS_i \mid i=0..n\}$ 中の，あるセグメント TS_n が，Why-Q に対する答えを含んでいるか否か判断を行うことである．このように Why 型の特徴を持った TS を **Why テキストセグメント(WTS)**，WTS ではない TS を **NotWhy テキストセグメント(NWTS)**と呼ぶ．

事実・結果（質問）セグメント特定は，判定された WTS 集合$\{WTS_i \mid i=0..n\}$において，質問に関する事

実・結果を含んでいる WTS_n が質問に対する回答を含んでいる可能性があると判断することである．

最後に，理由・原因となる回答抽出(Why-AE)は特定された WTS_n の中から Why-Q の答えとなる理由・原因を表す回答部分，Why-A を抽出する．

本論文では，文献[4]同様，事象や事柄に関する事実・結果を含む文を事実文，その回答となる理由・原因を含む文を，理由文と呼ぶ．

本研究の回答抽出も，文献[4]で述べられている回答位置の特徴を参考にする．文献[4]では，セグメント中の事実文起点として出現する理由語(から，からこそ等)と指示語(それ，以下，以上 等)の特徴語のパターンをもとに理由文の位置を前方(Case1)，後方(Case2,Case3)と事実文中(Case4)の特定を行っている．特徴語等の詳細は文献[4]を参照されたい．本研究では，ルールを緩和と簡潔化と目的として前方(Case1)，後方(Case2& 3)と事実文中(Case4) 3 つのケースを考慮する．

4. 提案手法

4.1. Grammar-gram と GrammarVerb-gram

形態素 Uni/Bi-gram を用いて学習を行なう手法[2]は，クローズドメインにおける Why-TSC と Why-AE に対しては効果的ではあるが，オープンドメインに対応したグローバルな学習器を作ることは困難である．

本研究では，TS 中の文法情報に着目して，より少ない情報でドメインに依存しないグローバルな学習器の構築を行うことで，Why-TSC と Why-AE を実行できる Grammar-gram と GrammarVerb-gram 方法を提案する．この手法では，ラベリングと学習を 1 度行なうだけで，オープンドメインでの Why-TSC と Why-AE が可能な学習器を構築できる．

日本語の単語は大きく分けて，内容語と機能語に分けられる．内容語は，名詞や動詞などの形態素からなり，文字通り，文の内容や動作などを表す語を示す．これに対して，機能語とは，助詞や助動詞など，文中において，内容語を繋げて文法の構造を構築する機能をもつ語を指す．

ここで，以下の２つのテキストセグメントは，機能語と内容語（下線太字部）に分けられており，文①は事実を表し，文②は理由を表しているのは明確である．

① 私は学生です．

② テストなので，学校に行った．

この文で，名詞の部分の“学生”や“学校”を，例えば“教授”や“塾”，に変えても依然として，①は事実を表し②は理由を表しており，提供される情報は変わるが述べられている形態（事実・理由），つまり“コンセプト”は全く変わらない．これは“～は～です”や“～なので～”といった機能語の部分が事実や理由を表している要素として判断が可能であるという事実に基づいている．つまり，あるセグメントが，WTS らしい要素を含むと判断するには，その文やセグメントが表す文法情報の役割が重要であると考えられる．

すべての形態素を考慮して TS を表したものを形態素 Uni-gram というのに対し，このように，TS の文法情報だけを用いて TS を表現する手法を Grammar-gram(G-gram)と呼ぶ．本稿では，基本的に機能語を Grammar として扱う．文法情報は，名詞などに比べてはるかに少なく，ドメインに依存しない特徴として TS を表現でき，Why-TSC と Why-AE のグローバルな学習器の作成が，可能である．

動詞も助詞などと同様，名詞に比べて，語彙数的にある程度普遍的であると考えられるので，G-gram に動詞を加えたものを GrammarVerb-gram(GV-gram)と呼び，GV-gram もグローバルな学習器作成のオプションとして考慮して，Why-TSC と Why-AE を行う．

G/GV/Uni-gram で扱う品詞の定義と，形態素解析されたセグメント“テスト/な/ので/学校/に/行っ/た”を例にして G/GV/Uni-gram 化したものを表 1に示す．

G-gram は，EDR 辞書の原因を表す文を抽出し，その文の機能語だけを抜き出して抽象化を行う手法[20]に類似している．しかし，文献[20]では，文の機能語だけを残した構造体を作成し，これから“ルールの抽出”を行ない，これを素性としてランキングしているのに対して，本研究では，機能語を含む品詞の集合を文法情報として G-gram を定義し，構造体ではなく，形態素レベルで，“ルールのマイニング”を行なった分類問題と見なし，グローバルな Why-QA の学習器を構築することを提案している．さらに，文献[20]は，GV-gram のような動詞は考慮しおらず，Why-AE にも言及されていない．

4.2. LogitBoost

本研究では，G/GV-gram を素性とした機械学習を行なうことにより，Why-TSC，Why-AE を分類問題と考えて，オープンドメインで識別が可能な学習器の自動構築を行う．本研究では，Why-TSC を，WTS クラスと NWTS クラスのバイナリークラスの分類，Why-AE を Case[1, 2&3, 4]と NoCase のマルチクラスへの分類ととらえる．学習には機械学習の 1 つである LogitBoost[18]を用いる．

LogitBoost は，AdaBoost[17]同様，Boosting を利用した機械学習の方法で，マルチクラス分類に対応している．LogitBoost は，クラスごとで弱識別器を用いて学習を行い事後分布が計算され，それを元に重みの更新を行っていく．その結果，J クラスの関数 $\{F_j(x)|j=1..J\}$ が作成され，分類は $\mathrm{argmax}_j F_j(x)$ として決定される．

LogitBoost を用いる理由は，速い学習スピードと高い識別能力を有する機械学習であるとともに，複雑な

表 1　G/GV/Uni-gram に使われる品詞と形態素例

Table. 1　POS used in G/GV/Uni-gram & morph example

N-gram	形態素例	品詞
G-gram	な，ので，に，た	助詞，助動詞，接頭詞，連体詞，接続詞
GV-gram	な，ので，に，た，行く	G-gram 品詞+動詞
Uni-gram	∀形態素	∀品詞

パラメータの調整の必要も無く，イタレーションの数だけを指定することで学習が開始でき，容易に高い判別結果が得られるからである．また，本研究では，回答位置抽出において，マルチクラス分類を想定しているので，従来のバイナリークラス分類器を用いた 1vsRest 手法より優れたマルチクラス分類を有する LogitBoost を，学習機構として用いることにした．

4.3. 学習データの収集方法

質問解析や文書解析といった部分の多くは，それだけで多くの研究課題を含み，また開発時間も要するので，本研究の目的を円滑に遂行し，より純度の高い学習器の構築するために，ドキュメントのセグメント化や Why-Q のキーワード化の処理はすでに行われているものと仮定，あるいは一部を手動で行うことにより，Why-TSC と Why-AE を評価することにした．

4.3.1. Why-QA 学習用 TS データの収集

Why-QA 学習用 TS データは，様々な Why-QA の言い回しを吸収するために Google を用いてインターネット上のドキュメントから，次の手順で TS の収集を行なう．1)キーワードを検索クエリとして，Google に検索をかける．2)Google の検索結果の上位 N 個のドキュメントから，事実文が含まれているドキュメントの判定を行う．ここで，事実文とはキーワードを全て含む 1 文とする．3)事実文が含まれる周辺の意味的にまとまった文を手動で判定し，TS データを収集した．

本研究では，文献[4]同様，質問文中のキーワード，キーワード+“なぜ”，キーワード+“どうして”の 3 つのクエリを使って Google で検索した．キーワードには文献[4]の付録 A.1 で使用している 10 質問と付録 A.2 の 5 質問を用いて，各検索結果の上位 100 ドキュメントから事実文周辺の，意味的にまとまった TS を抽出し，881 件の TS データを収集した．

4.3.2. TS データラベル付け

収集された TS データに対して，Why-TSC 用の学習のために，WTS と NWTS のラベル付けを手動で行なう．また，Why-AE 用のラベルとしては，WTS に対して，Case1，Case2&3，Case4 を判断してラベル付与を行い，NWTS は NoCase としてラベル付けを行う．

ラベル付けされたデータに対して，ChaSen[3]で形態素解析し，得られた形態素に対して，品詞情報も含むユニークな形態素の G-gram, GV-gram 素性として用いる．この素性をもとに，各 TS の G-gram，GV-gram 単位の出現頻度を求め，これを要素として各 TS を素性次元数でベクトル化を行う．これらのベクトル化されたデータを学習データセットとする．Uni-gram も，同様の手法で学習データセットを作成する．

表 2　素性数

Table. 2　Number of Attributes

	G-gram	*GV-gram*	*Uni-gram*
#Attrs	237	1118	3995

表 3　WTS/NWTS/Case のデータ分布

Table. 3　Data distribution for WTS/NWTS/Case

	WTS			NWTS
#Data	222			250/659
	Case1	*Case2&3*	*Case4*	*NoCase*
#Data	39	55	128	250/659

5. 評価実験と考察

5.1. 実験条件

本実験では，提案手法である G/GV-gram による Why-TSC と Why-AE の性能を，従来手法である Uni-gram を素性として用いる学習手法とルールベース手法の 2 つの手法と比較して評価を行う．

実験ではデータ数の偏りによる学習の精度低下を防ぐため，WTS の 222 個に対して，NWST から 250 個をランダムに選び，データ数の比率を同じ程度にして機械学習させ，実験を行った．G/GV/Uni-gram ごとのユニークな素性数を表 2に，各ラベル(クラス)のデータ数の分布を表 3に示す．

学習の際，データマイニングソフトウェア Weka[19]を使用した．Boosting の弱識別器には DecisionStump[3]を用いて 50，100，200，300，400，500 回のイタレーションで学習を行った．識別では 10 Folds Cross Validation (10FCV)を用いて，F 値を求めてその平均を識別結果とし，従来手法と比較して評価した．

5.2. Why テキストセグメント判定実験結果

Why-TSC の実験結果を表 4に示す．G-gram において，F 値が WTS 判定で 71.5%，NWTS 判定で 73.4%となり，ルールベースと比べて精度が向上した．GV-gram においても，F 値が WTS 判定で 72.4%，NWTS 判定で 74.5%と，ルールベースよりも高い F 値が得られた．

Uni-gram と比較すると，やや判定精度が劣るものの，G/GV-gram で使用した素性数はそれぞれ 237 個，1118 個であり，Uni-gram の約 4000 個に対して，G-gram では約 6%の素性数で約 70%の判定精度を有している．GV-gram も同様に，Uni-gram に比べて少ない素性で高い判定精度を有している．

[3] Decision stump は depth が 1 の決定木である．

表 4 Why-TSC F 値実験結果

Table. 4 Why-TSC F-Measurement Experiment Results

	G-gram	GV-gram	Morph-gram	RuleBased
F-Measure				
WTS	0.715	0.724	0.816	0.638
NTWS	0.734	0.745	0.844	0.672

表 5 Why-AE F 値実験結果

Table. 5 Why-AE F-Measurement Experiment Results

	G-gram		GV-gram		Morph-gram		RuleBased
	Expt1 *	Expt2**	Expt1	Expt2	Expt1	Expt2	
F-Measure							
Case1	0.444	0.368	0.493	0.325	0.454	0.450	0.242
Case2&3	0.494	0.372	0.485	0.360	0.507	0.422	0.304
Case4	0.794	0.521	0.818	0.515	0.803	0.652	0.521
NoCase	—	0.731	—	0.751	—	0.840	0.672

* -NWTS : ①の WTS の Case[1,2&3,4]のみを用いた分類　** +NWTS : ②の WTS の Case[1,2&3,4]と NWTS の NoCase を用いた分類

Uni-gram が G/GV-gram よりやや高い結果となったのは，Uni-gram を素性として用いた学習方法と 10FCV に関連性があると考える．10FCV は 1 つのデータセットを 10folds に分割し，9folds を学習に使い，残りの 1fold でテストを行うことを 10 回繰り返すことで評価を行なう方法である．データ数の少ない実験において有効な評価方法であるが，10FCV を用いて全ての形態素を用いた Uni-gram で学習を行なった場合，そのデータ特有の高い判別能力を持つ情報，つまり名詞が素性として働くので認識率が高くなると考えられる．いわば，Uni-gram+10FCV は 1 つのデータを 1 つのドメインと見なすことができ，そのドメインにおける評価とみなせる．逆に，G/GV-gram で用いた素性は文法情報で，一般性の高い情報なので，素性の一般性により，やや精度が劣るものと考えられる．

このことより，文法情報だけを学習に使用するという本手法は，名詞情報を含む Uni-gram よりも，一般性を持つ情報を学習し，グローバルな学習器を構成することができると考えられる．

5.3. 理由・原因となる回答抽出実験結果

Why-AE 実験評価は、2 つの方法を用いて行った．

① 正解セットに対するデータの分類度を評価するため，WTS データだけを用いて，Case[1,2&3,4]へのマルチクラス分類を行い，精度を評価する．

② WTS データと NWTS データの 4Case のデータを学習し、4 クラスへの分類を行い，精度を評価する．

表 5は，①(表には-NWTS と表記)，②(表には+NWTS と表記)の手法での G-gram，GV-gram による Why-AE と Uni-gram とルールベースの実験結果を示している．

実験②の NoCase を含めたマルチクラス分類における Why-AE では，ルールベースよりも相対的に高い結果となった．名詞情報を含む Uni-gram よりは劣るものの，5.2 と同様の理由で Uni-gram の精度が若干高いと考える．

実験①の結果は，正解セットの中でどれだけの精度で分類が可能であるかを調べたものである．この結果，②に比べて，高い精度で判別可能となっていることが分かる．特に Case4 は，F 値が G-gram で 79.4%，GV-gram では 81.8%と高い精度で回答抽出が可能となった．つまり，本手法は，Why-TSC の精度が高いほど，より正確な回答抽出が可能であると言える．

5.4. 学習速度の有効性の検証

Why-TSC と Why-AE の G-gram, GV-gram, Uni-gram, による学習の際に，50 と 500 のイタレーションに要した時間を比較する(Why-TSC/Why-AE と表記)．

Uni-gram では 98.1s-979.92s / 196.09s-2189.33s を要し，Iteration 数の増加に伴い，急激な学習時間の上昇が見られる．一方で，G/GV-gram では 5.25s-50.95s / 10.48s-104.74s と 27.24s-266.73s / 52.42s-531.91s と学習時間も短く，なだらかな学習時間の上昇となっており，Why-TSC と Why-AE の両方において，より高速に学習可能であることがわかる．Uni-gram 手法では，学習のデータ量が多くなれば素性数は急増するので，学習速度は格段に遅くなると考えられる．また，メモリー消費も増加する．さらに，G/GV-gram 手法は，新しい特徴のマイニングにおいても，素性の急激な増加を生じさせることなく再学習が可能であるため，未知のルールに簡単に対応できる．

以上より，ルールベース手法や Uni-gram 学習方法に比べ，G/GV-gram は，より効率的かつ実用的な手法であると言える．

5.5. 提案手法の応用例と問題点

本手法の利点は，ドメインに依存しない Why-QA システムを簡単に構築でき，QA システムの拡張を行なえることにある．例えば，既存の QA システムで用いている形態素解析ツールの形態素辞書において，本手法で文法と定めた品詞をもつ形態素全てを，機械学習用の素性として使用することにより，オープンドメイ