5Q-03

RITE-VALタスクを対象とした 表層類似度に基づくテキストの真偽判定

服部 昇平 † 佐藤 理史 † 松崎 拓也 †

† 名古屋大学大学院 工学研究科 電子情報システム専攻

1. はじめに

RITE[‡] (Recognizing Inference in TExt) とは，テキスト間で成立する含意，換言，矛盾の自動認識を目的とした評価型ワークショップである．これらの認識技術は，情報検索や質問応答，文書要約などの自然言語処理分野で広く共通する課題であり，その重要性は広く認識されている．

2014年に開催されたRITE-VAL [1] では，テキスト含意認識の応用課題として，テキストの真偽判定を題材としたタスク (FVサブタスク) が設定された．一般的な含意認識タスクでは，与えられたテキスト (**t**) と仮説 (**h**) に対して，**t** から **h** が推論可能であるかどうかを判定する．FVサブタスクでは，**t** の代わりに文書集合 **D** が与えられ，その **D** に基づき，**h** の真偽 (**Y**：**h** が真，**N**：**h** が偽) を判定する．なお，このタスクでは，大学入試センター試験の社会科の問題が用いられる．

本稿では，RITE-VALにおけるFVサブタスクを対象とした，表層類似度に基づくテキスト真偽判定システムついて報告する．本システムは，図1に示すように，含意認識とテキスト検索の2つのモジュールにより構成される．それぞれのモジュールは，以下の処理を行う．

テキスト検索 与えられた仮説 **h** の真偽を判定するための記述 (**t**) を文書集合 **D** から抽出する

含意認識 テキスト検索で得られた **t** が **h** を含意するかどうか判定する

以下では，まず，各モジュールについて概説する．続いて，FVサブタスクの結果を報告する．最後に，我々のシステムを用いた大学入試センター試験の自動解法について検討する．

2. 含意認識モジュール

含意認識モジュールには，RITE-2 [2] で用いた含意認識システム [3] を，一部改良したものを用いる．このシステムは，以下の2段階で含意認識を行う．

ステップ1 テキストペアの表層類似度を計算し，類似度が高い場合は **Y**，低い場合は **N** と判定する．

ステップ2 ステップ1で **Y** と判定した場合，テキストペアの差異を調査し，その結果を覆す必要があるかどうか判定する．

ステップ1では，表層類似度として，式 (1) で定義されるオーバーラップ率を用いる．

$$\mathrm{overlap\_ratio}(E;\mathbf{t},\mathbf{h}) = \frac{\sum_{x\in E} \min(\mathrm{f}(x,\mathbf{t}),\mathrm{f}(x,\mathbf{h}))}{\sum_{x\in E} \mathrm{f}(x,\mathbf{h})} \quad (1)$$

‡http://research.nii.ac.jp/ntcir/index-ja.html

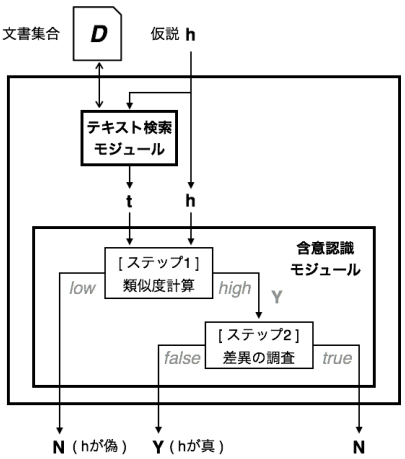

図 1: システム概要

表 1: 年代表現の正規化

タイプ	表現	正規化後
単年	n 年	n
世紀	n 世紀	$[n_{start} - n_{end}]$
範囲を表す	n_1 年から n_2 年	$[n_1\text{-}n_2]$
	n 年代	$[n_{start} - n_{end}]$
	n_1 世紀から n_2 世紀	$[n_{start} - n_{end}]$

ここで，$\mathrm{f}(x,\mathbf{t})$ は，集合 E の要素 x が，**t** 中に出現する回数を表す．なお，集合 E には，文字集合，形態素集合，文字 n-gram の集合などを想定する．どのような類似度がテキスト真偽判定に有効であるかは，今回新たに調査した (3章)．

ステップ2では，固有名詞と年代表現の不整合に着目する．含意関係にある2つのテキストでは，出現する固有名詞や年代表現が共通している場合が多い．そのため，「**h** に出現する固有名詞，または年代表現が **t** に出現しない場合，**t** から **h** への含意関係は成立しない」と判定する．

年代表現の照合では，単純な数字の比較だけでなく，「範囲」を表す表現への対処を新たに加えた．以下に例を示す．

> **t**：ジョン・ケネス・ガルブレイスは1943年から1948年にかけて「フォーチュン」誌の編集者を務めた．
> **h**：フォーチュンは，1940年代，ジョン・ケネス・ガルブレイスを編集員として起用した．

この例では，**h** に「1940年代」，**t** には「1943年から1948年」という表現が存在する．単純な数字の照合では，これらの表現が整合することを正しく判定できない．そこで，テキスト中の年代表現を表1に示す規則で正規化し，照合を行う．この場合，「1940年代」という表現は [1940-1949] に，「1943年から1948年」は [1943-1948] に正規化される．**h** から得られた [1940-1949] は，**t** から得られた [1943-1948] を包含するため，適切な照合が可能となる．

表 2: 分割単位とスコア計算方法に対する判定性能

分割単位	*E*	M-F1	th.
文 1-gram	*C* (文字集合)	55.7	0.90
文 2-gram		57.7	0.83
文 3-gram		57.2	0.89
段落	*C* (文字集合)	55.3	0.90
	K (漢字・カタカナ集合)	58.0	0.88
	W (形態素集合)	55.0	0.86
	N **(名詞集合)**	**59.7**	0.69

表 3: FV サブタスクにおけるシステム性能

データ	M-F1	Acc.	順位	
			run	team
開発用 (Y:210, N:300)	60.4	59.7	-	-
評価用 (Y:208, N:306)	59.5	57.3	6 / 30	2 / 9

3. テキスト検索モジュール

テキスト検索モジュールでは，与えられた仮説 **h** の正誤を判定するための記述 (**t**) を，文書集合 **D** から抽出する．ここでは，(1) どの単位 (e.g. 文，段落，節) で文書を分割するか，(2) 抽出する文書断片をどのように選択するか，が問題となる．

最初の問題に対しては，事前に付与された文章構造を用いる．FV サブタスクでは，教科書データが XML フォーマットで提供され，段落や節といった文書構造が明示されている．これらの中から最も適切な分割単位を選ぶ．

2 つめの問題に対しては，分割されたすべての文書断片にスコア付けを行い，その中から最もスコアが高いものを選択する．含意認識モジュールでの類似度計算と整合させるため，スコアには，式 (1) のオーバーラップ率を用いる．

分割単位やスコア計算には，いつくかの方法が考えられるため，最適な組み合わせを実験的に決定した．調査手順は，次のとおりである．まず，開発用データ中の **h** に対して **t** を抽出する．次に，得られた **t** のスコアと適当なしきい値を用いて **h** の真偽を判定し，性能を比較する．しきい値は，開発用データに対して M-F1[1] が最も高くなるように設定し，性能の比較は M-F1 で行う．

調査結果を表 2 に示す．候補には，表 2 に示す 7 通りの組み合わせを用いた．この結果より，分割単位として段落を，スコア (類似度) として名詞オーバーラップ率を採用する．

4. RITE-VAL フォーマルラン

FV サブタスクでは，タスクオーガナイザーより，2 種類の教科書データが提供された．本システムでは，これら 2 つのデータを文書集合 **D** として用いる．まず，それぞれの教科書データを用いて，独立に真偽判定を行う．その後，**Y** に対する論理和をとることで最終的な出力を決定する．

表 3 に，開発用データと評価用データのそれぞれに対する正解率 (Acc.) と *MacroF*1(M-F1) を示す．この表の順位は，フォーマルランにおける順位を意味する[§]．この表に示すように，我々のシステムは，30 手法中 6 位，9 チーム中 2 位の成績であった．

[§]システム性能は M-F1 によって評価される．

表 4: センター試験 (世界史 B) 結果

年度	FV 形式 [点]	空欄形式 [点]	計 [点]
2005	31 / 70 (44%)	6 / 25 (24%)	37 / 95 (39%)
2007	27 / 67 (40%)	15 / 15 (100%)	42 / 82 (51%)
2009	40 / 82 (49%)	3 / 6 (50%)	43 / 88 (49%)
2011	40 / 79 (51%)	6 / 6 (100%)	46 / 85 (54%)
2013	22 / 68 (32%)	5 / 8 (63%)	27 / 76 (36%)

5. 大学入試センター試験 (世界史 B)

作成したテキスト真偽判定システムを用いて，実際の大学入試センター試験 (世界史 B) の自動解答を試みた．FV サブタスクのデータもセンター試験の問題で構成されているが，テキストの一部は人手で加工されている．そのため，元の試験問題をそのまま用いた場合，どの程度の正解率が得られるか調査した．なお，試験問題データには,「ロボットは東大に入れるか。」[¶]プロジェクトで作成された XML データを用いた．

今回は，以下の 2 種類の設問を解答対象とした．

FV 形式 選択肢の正誤判定により解答を選択できる設問

(e.g. 正しいものを，次の①～④のうちから一つ選べ。)

空欄形式 問題文中の空欄に入る適切な語句を選ぶ設問

FV 形式の設問では，それぞれの選択肢の真偽を判定することで解答を選択する．複数の選択肢が **Y** と判定される場合は，最もスコアが高いものを選択する．

空欄形式の設問では，以下の手順で解答を選択する．

1. 選択肢の語を含む段落を教科書から抽出
2. 空欄を含む 1 文に選択肢の語を代入し，1 で得られたすべての段落との類似度 (スコア) を計算 (その中の最大値をその選択肢のスコアとする)
3. スコアが最も高い選択肢を解答する

いずれの形式においても，スコアには，式 (1) による名詞オーバーラップ率を用いる．

2005 年から 2013 年における奇数年度の試験問題 (5 年分) に対する結果を表 4 に示す．年度によって多少の差は見られるが，平均して 45%程度の正解率が得られた．4 択問題としてのチャンスレベル (25%) は大きく超えているが，人間と比較すると高い正解率とは言い難い．センター試験では，問題文中の下線部など，問題を解くために参照すべき箇所が多く存在する．現在の解法では，選択肢のテキストしか用いていないため，問題文中の情報を上手く活用する必要がある．

参考文献

[1] Suguru Matsuyoshi, Yusuke Miyao, Tomohide Shibata, Chuan-Jie Lin, Cheng-Wei Shih, Yotaro Watanabe, and Teruko Mitamura. Overview of the NTCIR-11 Recognizing Inference in TExt and Validation (RITE-VAL) Task. In *Proceedings of the 11th NTCIR Conference* pp. 223–232, 2014.

[2] Yotaro Watanabe, Yusuke Miyao, Junta Mizuno, Tomohide Shibata, Hiroshi Kanayama, C.-W. Lee, C.-J. Lin, Shuming Shi, Teruko Mitamura, Noriko Kando, Hideki Shima, and Kohichi Takeda. Overview of the Recognizing Inference in Text (RITE-2) at the NTCIR-10 Workshop. In *Proceedings of the 11th NTCIR Conference*, pp. 385–404, 2013.

[3] 服部昇平, 佐藤理史, 駒谷和範. 表層類似度に基づく日本語テキスト含意認識. 人工知能学会論文誌, Vol. 29, No. 4, pp. 416–426, 2014.

[¶]http://21robot.org/