2006－NL－174（6）
2006／7／27

意味的等価性検証に基づく記述式解答文の採点法

秋葉 泰弘† 田中 貴秋† 須山 敬之† 永田 昌明†

† NTT コミュニケーション科学基礎研究所
〒 619-0237 京都府「けいはんな学研都市」光台 2-4
E-mail: †{akiba,takaaki,suyama}@cslab.kecl.ntt.co.jp, ††nagata.masaaki@lab.ntt.co.jp

あらまし 本稿では，e-Learning における記述式解答文の採点自動化に向け，文同士の意味的等価性を検証すると言う問題を取り上げる．採点の自動化が取り組まれてきた記述式解答文は，エッセイや小論文といった比較的長い文章である．エッセイ評価では文章の内容を評価するために，文書検索と同様，採点対象の文章を頻度や TF・IDF の単語ベクトル (bag-of-words) にエンコードし，模範解答や採点済み被験者解答とのベクトル間類似度を計算してきた．文や文書の意味内容をより詳細に比較・検証するためには，語と語の意味的関係を捉えた構文意味情報を活用し得る．本稿ではまず，先駆的に取り組まれてきた自動採点技術として，エッセイ評価技術を概観する．また，自然言語処理の様々な分野で研究されてきた文や文書の意味内容の等価性を検証する関連技術を振り返り，それらの問題点について述べる．次に構文意味情報に基づき文同士の意味的等価性を検証する手法を提案する．最後に，今後予定している提案法の評価方法および評価に向けて準備中の評価データについて述べる．

キーワード 自動採点，記述式解答，依存構造，二項関係，同義語

Grading Examinee's Answer Sentences by Verifying Syntactic and Semantic Compatibility

Yasuhiro AKIBA†, Takaaki TANAKA†, Takayuki SUYAMA†, and Masaaki NAGATA†

† NTT Communication Science Laboratories
2-4, Hikaridai, "Kansai Science City", Kyoto 619-0237 Japan
E-mail: †{akiba,takaaki,suyama}@cslab.kecl.ntt.co.jp, ††nagata.masaaki@lab.ntt.co.jp

Abstract This paper addresses the problem of automatically verifying whether contents of two sentences are semantically equivalent to each other. Most existing methods classified with Essay Testing such as E-rater, which verify the content equivalency of two essays by using similarity measure based on bag-of-words. To realize more advanced content equivalency verification, it is possible to utilize semantic and/or syntactic information. The authors propose a new verification method based on semantic and/or syntactic information extracted from dependency structures of verification target sentences.

Key words Automatic Rating, Descriptive Test Answer, Dependency Structure, Binary Relation, Synonym

1. はじめに

本稿では，e-Learning における記述式解答文の採点自動化に向け，文同士の意味的等価性を検証すると言う問題を取り上げる．インターネットの普及に伴い，Web をインターフェースとした e-Learning システムが企業内研修や語学学校向けに開発され，学校へ通わなくても遠隔地で研修や教育が受けられるようになってきた．e-Learning システムの重要な機能の 1 つに習熟度測定がある．習熟度測定では学習内容がどの程度習熟したかを計測する．紙ベースの試験の解答形式には選択式解答と記述式解答があるが，採点が自動化されたいるのは選択式解答のみであり，記述式解答の採点自動化が望まれている．

記述式試験問題は次の 3 つに大別できる．一番目は専門用語の穴埋め問題で，専門用語を正しく覚えているかを問う．穴埋め問題の採点は表層レベルの文字列比較により自動化が可能である．二番目はもう少し長い表現の記述（名詞句～数文）を求める問題で，学習テーマに関する知識を問う．この手の問題では，一字一句模範解答と表現が一致していなくとも，意味内容が正しければ正解として採点する必要がある．この種の問題に対する採点の自動化は研究されて来なかった．構文意味情報の

レベルで照合する技術の更なる進展が必要である．三番目は小論文やエッセイ等の長い記述を求める問題で，被験者の記述能力全般を問う．技術的にはエッセイ評価 (Essay Testing) [1]～[5] としてかなり研究が進んでおり，実用的なレベルに達している．例えば，アメリカの経済大学院の入試試験である GMAT (Graduate Management Admission Test) のある部分を採点では，**E-rater** (Electronic Essay Rater) [1] と言う自動採点技術が採用されている．

E-rater に代表的されるエッセイ評価技術の多くでは，文書の意味内容を評価するために，文書検索と同様，採点対象の文章を頻度や TF・IDF の単語ベクトル (bag-of-words) にエンコードし，模範解答や採点済み被験者解答とのベクトル間類似度を計算する．

bag-of-words は文を構成する内容語が一致してるが意味内容が異なるような文同士を区別することができない．例えば，以下の文 1) と 2) は意味内容は異なるが，以下のように，同じ単語ベクトルにエンコードされてしまう．この手の問題を対処し，文や文書の意味内容をより詳細に比較・検証するためには，語と語の意味的関係を捉えた構文意味情報を活用することが考えられる．

1)	太郎は花子が好きだ．
2)	花子は太郎が好きだ．

n	(	太郎	花子	好き	)
1)	(	1	1	1	)
2)	(	1	1	1	)

本稿では以下，2. 節で，先駆的に取り組まれてきた自動採点技術として，エッセイ評価技術を概観する．また，自然言語処理の様々な分野で研究されてきた文や文書の意味内容の等価性を検証する関連技術を振り返り，それらの問題点について述べる．次に 3. 節で，構文意味情報に基づき文同士の意味的等価性を検証する手法を提案する．4. 節で，今後予定している提案法の実験評価および評価に向けて準備中の評価データについて説明する．最後 5. 節で本稿をまとめる．

2. 関連研究

2.1 エッセイ評価

本節では，記述式解答の自動採点の中で研究事例が多いエッセイ評価について概観し，それらの問題点について述べる．代表なエッセイ評価技術には，1) E-rater (Electronic Essay Rater) [1]，2) IEA (Intelligent Essay Assessor) [3]，3) Jess (Automated Japanese Essay Scoring System) [4], [5] がある．

E-rater [1] は，エッセイ評価技術のなかでも特に有名な自動採点法であり，1. 節で述べたように，アメリカの有名な試験 GMAT で運用されている．E-rater の特徴には，1) 文書に表れる文構造の多様性，2) 文同士の論理的な接続関係，3) 論述のテーマに関連した語彙の選択，以上の３つの視点でエッセイを評価する点にある．これらを計量するために，色々な観点で文書の特徴量を計測する．例えば，文構造の多様性を測るために文構造の種別毎の集計したり，接続詞などの文同士の論理的な関係を示す手がかり語数を数えたり，文の記述内容を検証するために文書を用語の単語頻度ベクトルで表し，既に採点済みの解答とのコサイン類似度を計測したりする．採点済みの文書をこれらの特徴量でエンコードし，多変量解析の重回帰分析を行なう．新たな文書を採点する際には，重回帰分析によって学習された回帰直線を用いる．

二番目の代表例 **IEA** (Intelligent Essay Assessor) [3] は，E-rater の特徴 3) である語彙選択の適切性に関して評価を行うよう設計されている．基本的なアイデアは，文書検索の LSI (Latent Semantics Indexing) のように，採点済みの文書を構成単語の頻度や TF*IDF 値を成分とする単語ベクトル（次元は t とする）で表現し，行ベクトルが文書の単語ベクトルに対応するように文書群に対応する $t \times d$ の行列（d は文書数）を作成し，この行列を多変量解析の特異値分析（主成分分析）に掛け，得られた固有ベクトルを基底するベクトル空間上に各文書をエンコードし直す．新たな文書を採点する際には，この新しい基底で張られるベクトル空間上で，採点済みの文書とのコサイン類似度を計算し，最もコサイン類似度の高い採点済み文書の得点をこの新たな文書の得点とする．

三番目の代表例 **Jess** (Automated Japanese Essay Scoring System) [4], [5] は，簡単に言うと E-rater の日本語版である．E-rater との違いは内容評価法が IEA の内容評価法で置き換えられている点である．

上記の説明から判るように，これらの３つのエッセイ採点技術における内容評価はいずれも，単語ベクトル同士 (bag-of-words) のコサイン類似度に基づいている．1. 節の例文 1) と 2) で例示したように，内容語が一致してるが意味内容が異なるような文同士を区別することができない．この手の問題を対処し，文や文書の意味内容をより詳細に比較・検証するためには，語と語の意味的関係を捉えた構文意味情報を活用することが考えられる．そこで本稿では構文意味情報に基づき文同士の意味的等価性を検証する手法を提案する．

2.2 内容等価性判定

本稿では文同士の意味的等価性を検証すると言う問題を取り上げている．この問題に関連する技術としては，文対応付け技術，翻訳自動評価，およびカーネル法が考えられる．以下本節ではこれらの技術を振り返り，これらの問題点について述べる．

2.2.1 文対応付け

統計翻訳システムの構築法は，人手で翻訳ルールを記述するというアプローチから，対訳コーパスから統計モデル（翻訳モデルと言語モデル）を学習するというアプローチへ大きく変遷し，近年の目覚しい進歩を遂げている．対訳コーパスを簡便に構築するために，第一言語の文を第二言語の文に対応させる文対応付け技術が盛んに研究されている．文対応付け技術では，第一言語の文と第二言語の文が意味的に等価である文対を見つけるタスクであり，その意味で本稿で取り上げている問題と関連深い．

文対応付けにおける等価性を検証する指標としては，文長 [6], [7] や対応する訳語対の数 [8] が用いられている．文長は

文字数[6]や数単語[7]といった表現単位に基づき計測される．文対応付け以外のタスク中で，任意の文同士を文長の一致だけで意味的に等価であると判断するのはかなり無理がある．特に，何文字以内で記述せよと言ったタイプの設問の場合には，解答の記述はほぼ同じ長さで記述されていると言ってよく，如何なる解答も模範解答と文長がほぼ一致しており，正解であると判断されてしまう．

対応する訳語対の数を指標とした文対応付け[8]は，対応付けに用いる辞書を対訳辞書から同義語辞書に置き換えれば，原理的には本稿で取り上げている問題に適用可能であるが，bag-of-word が bag-of-synset に置換わっただけなので，単語ベクトル同士 (bag-of-words) のコサイン類似度の問題点をそのまま内包していると言える．

2.2.2 翻訳自動評価

統計的機械翻訳では対訳コーパスからの学習を行うため，何らかの目標関数を定めるとその目標関数の意味で統計モデルを最適化することが可能である．この最適化の目標関数として統計的機械翻訳では，参照訳（理想訳）と機械訳を N-gram レベルで照合した際の N-gram の平均共有率[9],[10]が用いられる．

この指標は単語のローカルな共起情報が類似性している文同士は意味的に等価であるという仮説に基づいて，訳文同士の意味内容を検証している．動詞とその動詞に係る名詞の対のように比較的離れた共起情報により規定される意味内容については原理的に考慮できない．

2.2.3 カーネル法

機械翻訳以外の言語処理分野においても，学習に基づく解析器が様々提案されている．ここで特に注目したいのはカーネル法とよばれる類似度を測る技術で，隣接しない単語同士の依存関係を考慮できるストリングカーネル (String Kernel)[11],[12]や依存構造等の木構造で規定される多項関係を考慮できるツリーカーネル (Tree Kernel)[13]〜[17]である．

意味的等価性を検証する指標としてこれらストリングカーネルやツリーカーネルを用いた場合には，次のような問題点が考えられる．ストリングカーネルは全ての部分列が検証対象の2文に含まれるか否かを原理的に考慮するため，ベクトル成分の多くは依存関係の無い語群で構成される部分列の有無に対応する可能性が高くなる．ストリングカーネルで定まるベクトル同士のコサイン類似度を指標に用いた場合，意味的等価性な二文と等価でない二文をそれぞれコサイン類似度に値の大きな差がなくなり，分別性能が低くなる恐れがある．ちなみに，カーネルは通常サポートベクターマシンの枠組みで用いるが，ベクトルの各成分に学習により定まる重み値が乗算されるため，このような問題は起き難いと考えられる．

ツリーカーネルを意味的等価性検証に用いた場合には，上記のカーネルの問題点に加え，比較する木構造がどの程度正しいかが懸念される．例えば，意味的等価性を検証する二文をツリーカーネルを用いて依存構造同士のコサイン類似度を計測する状況を想定したとする．依存構造解析の精度は現在，係り受けレベルでは約 90%であるが，文レベルでは約 50%に留まる．ツリーカーネルでは共通する部分木の数で文同士の類似度を計測するため，ノード数の少ない部分木の部分木レベルの解析精度は高いことが期待されるが，ノード数の多い部分木では部分木レベルの解析精度は低くなる恐れがある．従って，大きな部分木に対応するベクトル成分が多いほど，コサイン類似度も信頼できなくなる．

3. 提案手法

本稿で取り上げた文同士の意味的等価性を検証すると言う問題に対して，構文意味情報に基づき文同士の意味的等価性を検証する手法を提案する．

2. 節で上げた関連技術の様々な問題点を考慮すると，意味的等価性を検証する手法は以下の用件を満足することが望まれる．

- 語と語の意味的関係を捉えるために構文意味情報を活用する．
- 大きな依存構造に対応する構文意味情報は利用を控える．
- 依存構造中のノードを照合する際には，語彙レベルの意味の類似性を考慮する．

提案手法の処理手順は以下の通り．

(Step 1) 意味的等価性を検証する二文 $S1$，$S2$ を形態素解析器および依存構造器にかけ，該二文を依存構造に変換する．以下，$S1$ の依存構造を $D1$ と表記する．$S2$ についても同様．

(Step 2) 依存構造 $D1$ に含まれる依存関係の二項関係全てを抽出する．以下，$D1$ から抽出した二項関係の集合を $B1$ と表記する．$D2$ についても同様．

(Step 3) 集合 $B1$ の要素と $B2$ の要素の全ての組合せに対して，二項関係同士の照合を行う．二項関係のノードは自立語同士が一致または互いに同義語である場合に限り，照合されたとする．照合された場合の数の総和を求める．この総和を $K(B1,B2)$ と表記する．

(Step 4) $K(B1,B2)$ を正規化した $N(B1,B2)$ を求める．即ち，$N(B1,B2)=K(B1,B2)/(\sqrt{K(B1,B1)}\sqrt{K(B2,B2)})$．ここで，K(B1,B1) は集合 B1 の要素数を表記する．K(B2,B2) も同様．

(Step 5) $N(B1,B2)$ が閾値 α より以上であれば，$S1$ と $S2$ の意味的内容は等価であると判断し，$N(B1,B2)$ が閾値 β 以下であれば，$S1$ と $S2$ の意味的内容は等価でないと判断する．それ以外であれば，$S1$ と $S2$ の意味的内容は等価性は不明であると判断する．なお，閾値 α,β は，訓練事例により予め閾値学習を行ない予め推定しておく．

4. 実験評価

本節では以下，今後予定している提案法の評価方法および評価に向けて準備中の評価データについて述べる．実験結果については別途報告する．

4.1 実験方法

記述式問題の被験者解答を人間の採点者に予め採点して貰い，被験者解答に○×△の三段階のいずれかの評点を割振る．

下記の評価法 1 および評価法 2 を評価基準にして，指標を 1) 提案法，2)Bag-of-Words，3) 文節のストリングカーネル，4) 依存構造木のツリーカーネルに切替えて，性能を比較する．

評価法 1 指標と評点との相関係数を比較

評価法 2 指標により閾値判定を行ない，判定の正解率を比較

なお，これから予定している提案手法のを実験評価では，(Step1) で用いる形態素解析器と依存構造解析器としては，Mecab [18] と Cabosh [19] を用いる．

4.2 評価データ

提案手法を実験評価するために評価データを現在準備している．準備している評価データの概要は以下の通り．

記述式試験問題として情報処理資格試験に対する模範解答と被験者解答を集め，被験者解答については人間の採点者に採点してもらう．模範解答については複数通りの解答があり得る場合には，別解を含めて収集する．

試験問題，解説，模範解答，採点済み人手解答に形態素タグや語義情報タグを付与する．我々の研究グループではこれまで Lexeed [20]〜[22] と言う基本語彙に対する語義データベースを構築すると共に，Lexeed の語義情報を付与した語義タグ付きコーパス (Sense Bank) [23] を構築してきた．将来的には，Sense Bank から学習した語義タガーを構築する予定である．そのため，語義タガーが想定する形態素タグや語義情報タグの正解をも人手で付与している．

提案法の (Step 3) で用いる同義語集合としては，Lexeed の各語義を定義する語義文から自動的に抽出した同義語集合 [24]〜[26] を人手でチェックし，誤りを修正した同義語集合を用いる．

5. おわりに

本稿では，e-Learning における記述式解答文の採点自動化に向け，文同士の意味的等価性を検証すると言う問題を取り上げた．この問題に関連する技術として，エッセイ評価，対訳コーパスを構築するための文対応付け，翻訳自動評価，サポートベクターマシンにおけるカーネル法を振り返り，これらの関連技術を意味的等価性を検証に適用した場合の問題点について説明した．問題点を回避するような意味的等価性検証技術として，依存構造から抽出した 2 項関係で規定される構文意味情報に基づく手法を提案した．最後に今後予定している提案法の評価方法および評価に向けて準備中の評価データについて述べた．実験結果については別途報告する．

文　献

[1] J. Burstein, K. Kukich, S. Wolff, C. Lu, M. Chodorow, L. Braden-Harder and M. D. Harris: "Automated scoring using a hybrid feature identification technique", Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics ACL-COLING-1998, pp. 206–210 (1998).

[2] E. B. Page, J. P. Poggio and T. Z. Keith: "Computer analysis of student essays: Finding trait differences in the student profile", Proceedings of AERA/NCME Symposium on Grading Essays by Computer (1997).

[3] P. W. Foltz, D. Laham and T. K. Landauer: "Automated essay scoring: Applications to educational technology", Proceedings of World Conference on Education Multimedia, Hypermedia and Telecommunications: EdMedia-1999 (1999).

[4] 石岡, 亀田："コンピュータによる小論文の自動採点システム jess の試作", **16**, 1, pp. 3–18 (2003).

[5] T. Ishioka and M. Kameda: "Automated japanese essay scoring system based on articles written by experts", Proceedings of COLING-ACL-2006 (2006).

[6] W. A. Gale and K. W. Church: "A program for aligning sentences in bilingual corpora", Computational Linguistics, **19**, 1, pp. 75–102 (1993).

[7] P. F. Brown, J. C. Lai and R. L. Mercer: "Aligning sentences in parallel corpora", 29th Annual Meeting of the Association for ational Linguistics: ACL-1991, pp. 169–176 (1991).

[8] M. Kay and M. Roscheisen: "Text-translation alignment", Computational Linguistics, **19**, 1, pp. 121–142 (1993).

[9] K. A. Papineni, S. Roukos, T. Ward and W.-J. Zhu: "Bleu: a method for automatic evaluation of machine translation", Technical Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, NY, pp. 257–258 (2001).

[10] G. Doddington: "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics", Proceedings of ARPA Workshop on Human Language Technology, San Diego, California, pp. 257–258 (2002).

[11] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini and C. Watkins: "Text classification using string kernel", Journal Machine Learning Research, **2**, pp. 419–444 (2002).

[12] N. Cancedda, E. Gaussier, C. Goutte and J.-M. Renders: "Word-sequence kernels", Journal Machine Learning Research, **3**, pp. 1059–1082 (2003).

[13] M. Collins and N. Duffy: "Convolution kernels for natural language", Proceeding of Neural Information Processing Systems (NIPS-2001) (2001).

[14] M. Collins and N. Duffy: "Parsing with a single neuron: Convolution kernels for natural language problems", Technical Report, University of California at Santa Cruz (2001).

[15] 鈴木, 佐々木, 前田："階層非循環有向グラフカーネル", 電子情報通信学会論文誌, **J88-D-II**, 2, pp. 230–240 (2005).

[16] 鹿島, 坂本, 小柳："木構造データに対するカーネル関数の設計と解析", 人工知能学会論文誌, **21**, 7, pp. 113–121 (2006).

[17] 平尾, 鈴木, 磯崎, 前田："単一言語コーパスにおける文の自動対応付け手法", 情報処理学会論文誌, **46**, 10, pp. 2533–2545 (2005).

[18] 工藤, 松本："チャンキングの段階適用による係り受け解析", 情報処理学会論文誌, **43**, 6, pp. 1834–1842 (2002).

[19] 工藤："形態素周辺確率を用いた分かち書きの一般化とその応用", 言語処理学会年次大会発表論文集, C3-3 (2005).

[20] F. Bond, 藤田, 田中, 中岩："日本語の統語・意味コーパス「檜」", 言語処理学会年次大会発表論文集, S2-3 (2005).

[21] F. Bond, S. Fujita, C. Hashimoto, K. Kasahara, S. Nariyama, E. Nichols, A. Ohtani, T. Tanaka and S. Amano: "The hinoki treebank: Working toward text understanding", Proceedings of COLING 2004 5th International Workshop on Linguistically Interpreted Corpora, pp. 7–10 (2004).

[22] F. Bond, S. Fujita, C. Hashimoto, K. Kasahara, S. Nariyama, E. Nichols, A. Ohtani, T. Tanaka and S. Amano: "The hinoki treebank: A treebank for text understanding", Proceedings of the First International Joint Conference on Natural Language Processing (IJCNLP-2004), pp. 554–559 (2004).

[23] T. Tanaka, F. Bond and S. Fujita: "The hinoki sensebank — a large-scale word sense tagged corpus of japanese — ", Proceedings of Workshop on Frontiers in Linguistically Annotated Corpora 2006, A Merged Workshop with 7th International Workshop on Linguistically Interpreted Corpora (LINC-2006) and Frontiers in Corpus Annotation III, COLING/ACL 2006 (2006).

[24] E. Nichols, F. Bond and D. Flickinger: "Robust ontology acquisition from machine-readable dictionaries", Proceedings

of 19th International Joint Conference on Artificial Intelligence (IJCAI-2005), pp. 1111–1116 (2005).

[25] F. Bond, 藤田, 橋本, 笠原, 成山, E. Nichols, 大谷, 田中, 天野: "日本語ツリーバンク「檜」: 言語理解のためのコーパス", 情報処理学会 研究報告「自然言語処理」, 第 2004-NL-159 巻, pp. 83–90 (2004).

[26] F. Bond, E. Nichols, S. Fujita and T. Tanaka: "Acquiring an ontology for a fundamental vocabulary", Proceedings of 20th International Conference on Computational Linguistics (COLING-2004), pp. 1319–1325 (2004).

[27] T. Tanaka, F. Bond, S. Oepen and S. Fujita: "High precision treebanking blazing useful trees using pos information", Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-05), pp. 330–337 (2005).

[28] F. Bond, 藤田, 橋本, 成山, E. Nichols, 大谷, 田中: "精細な文法に基づいたツリーバンク「檜」の構築", 情報処理学会 研究報告「自然言語処理, 第 2004-NL-159 巻, pp. 91–98 (2004).

[29] T. Tanaka, F. Bond, S. Oepen and S. Fujita: "High precision treebanking in the hinoki project", 言語処理学会年次大会発表論文集, A5-6 (2005).

[30] R. Dridan and F. Bond: "Sentence comparison using robust minimal recursion semantics and an ontology", Proceedings of COLING-ACL-2006 Workshop on Linguistic Distances (2006).

[31] H. Kashima and A. Inokuchi: "Kernels for graph classification", Proceedings of 1st ICDM Workshop on Active Mining: AM-2002 (2002).

[32] 松本, 北内, 山下, 平野, 松田, 高岡, 浅原: "形態素解析システム『茶筌』version 2.3.3 使用説明書" (2003).