英作文の誤り分類の推定

フラナガン，ブレンダン
九州大学大学院システム情報科学府情報知能工学専攻

殷，成久
九州大学情報基盤研究開発センター

鈴木，孝彦
九州大学情報基盤研究開発センター

廣川，佐千男
九州大学情報基盤研究開発センター

出版情報：情報処理学会研究報告. 2013 (A-4-1), pp.1-6, 2013-03-14. 情報処理学会九州支部
バージョン：

英作文の誤り分類の推定

フラナガン・ブレンダン[†1,a]　殷成久[†2,b]　鈴木孝彦[†2,c]　廣川佐千男[†2,d]

学習者の特性を把握し、それに応じた指導をすることは、教育システムの重要な課題である。本論文では、英作文の学習について、学習者の誤り特性を抽出するため、学習者の作文の誤りの種類(Kroll 1990, Weltig 2004)を推定するシステムを構築し、その性能を評価した。具体的には外国語学習コミュニティ Lang-8 に書かれた英文日記のデータに機械学習を適用した。

Error Classification of English Writing Lessons

BENDAN FLANAGAN[†1,a]　CHENGJIU YIN[†2,b]　TAKAHIKO SUZUKI[†2,c]
SACHIO HIROKAWA[†2,d]

An important issue in education systems is the ability to determine the characteristics of a learner and then to provide suitable guidance in response. In this paper a system was built to classify a learner's errors into categories (Kroll 1990, Weltig 2004) with the purpose of identifying the error characteristics of a learner studying English composition. More specifically, machine learning was undertaken on the writings of learners on the Lang-8 foreign language learning community site.

1. はじめに

近年、決まった教室内にとらわれない、Web 上での語学学習が盛んに行われている。特に、異なる母国語を持つ者同士が､お互いに母国語を教えあうサイトが流行している。例えば、日本人学生 A が英語で書いた文章を Web に掲載し、英語を母国語とする B がその文章を添削することができる。逆に、B が日本 語で書いた文章を A が添削するといった相互学習も可能である。Lang-8[*1] は代表的な外国語の相互添削を支援するサイトである。

相互添削を支援するサイトには、数多くの添削データが保存されており、これを活用できれば、学習の効果は一層高まると考えられる。我々は Lang- 8 のデータを用い、手作業で学生の誤りパターン（Error Pattern）を分類し、誤りパターン別のクイズシステムを構築した[8]。このシステムを利用して、学生の弱点（学生特有の誤りパターン）を把握し、その弱点に対して繰り返し学習を行うことで学習効果が上がることを示した。しかし、手作業で誤りパターンを分類するには、時間と労力が必要である。

近年の機械学習理論の研究の進歩はめざましいものがある。特に、サポート・ベクター・マシン(SVM) は固定次元ベクトルの効果的な分類のための標準的な技術となっている。SVM は識別性能が高く、多くの分野で利用されている[5]。そこで、本稿では SVM を用いて、 誤り分類推定を行う。

我々は、学習者の誤り特性を抽出することを長期的な目的としている。本稿では、 Lang-8 に書かれた英文日記から、まず､誤りが訂正された 500 件の文をランダムに 選んだ。次に、これらの文について、Kroll[6], Weltig[7]の作文誤り分類のどれに該当するか英語を母国語とする第一著者が識別を行った。Lang-8 では､原文と添削文の両方があり、添削文書には、削除や挿入などの訂正情報タグが付加されていることもあるが、必ずしも全ての添削結果に詳細なタグが書かれているとは限らない。そこで筆者らは、原文と添削文のアラインメントにより、削除と挿入単語を機械的に抽出した。こうして得られる削除や挿入タグ付きの単語も、一般の単語と同様に、誤り分類の推定に利用した。こうして得られた学習データに対しパターン分類器 SVM を適用して、 誤り分類の推定を行い、その推定性能を評価した。

2. 関連研究

これまで、外国語ライティング（記述）の学習者の行動についての実証的な研究は、主に閉じた教室内で行われてきた。学習環境を教室内に限定することで、ライティング内容に影響を及ぼすと思われる、ライティングの主題、ライティングが行われる環境、記述者の状況等の要因のコントロールが可能になった。

Kroll[6]は、時間に制限のある教室内と、時間に余裕がありかつプレッシャー の少ないと予測される自宅との二種類の環境で、学生のライティングにどのような差が出るかを比較実験している。学生のライティング中にあらわれる

†1 九州大学大学院システム情報科学府 情報知能工学専攻
Graduate School of Information & Research Institute for Science and Electrical Engineering, & Information Technology, Kyushu University, 6-10-1, Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan
†2 九州大学　情報基盤研究開発センター
Research Institute for Information, Kyushu University, 6-10-1, Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan
a) bflanagan.kyudai@gmail.com
b) yin@cc.kyushu-u.ac.jp
c) suzuki@cc.kyushu-u.ac.jp
d) hirokawa@cc.kyushu-u.ac.jp
*1 http://lang-8.com/

誤りを分類し、出現頻度を測定することで、二種類の環境におけるライティング内容の比較を行っていた。誤りの分類については英語教師が手作業で行っている。

Weltig[7]は、英語教師が外国語学習者のライティングを採点する際に、異なる種類の誤りが点数にどのような影響を与えるかを調査している。誤りの分類としては Kroll[6]の分類と類似したものを採用している。ある種類の誤りの頻度が、他の種類の誤りよりも、採点に大きな影響を与えることを報告された。

SVM その他の機械学習アルゴリズムを使用した、英文中の 誤り推定が研究されている。平野ら[2]は、英語のテクニカルペーパー中の冠詞誤りを検出するために検索エンジンの検索結果を利用した。彼らは文を構文解析し、タグ付した後、文の構造に基づいて検索クエリーを生成した。検索結果のヒット数から、入力文が誤りを含んでいるか否かを判定している。谷本ら[4]は英文中の単語誤りを同定するための指標として、検索結果数を利用する方法を調査した。彼らは、学習データとして NICE (Nagoya Interlanguage Corpus of English) 中の３-グラム、および 4-グラムを使用して、英文中に誤りが含まれているか否かを判定するためのモデルを作成している。

また、質問の分類やフォーマルな学術論文中の英語の質の 評価に注目する研究もある。鈴木ら[1]は質問文を分類する目的で、n-グラムと SVM を使うことを検討した。彼らはまず、質問タイプを分類識別する際に有用な特徴となる n-グラムを発見する方法を提案し、次いで、それらの特徴を SVM 学習データとして用い、質問クラス分けモデルを生成した。10,000 のサンプル質問に対してこのモデルを適用した結果、従来の方式と比較して優れているという結果を得ている。

Zhang[9]ではどの機械学習の手法が誤りの分類に有効か分析した。 彼らは TREC 英語コーパスについて単語、単語 N グラム、構文木をデータとして 機械学習を行った。その結果、テキストの表面的な特徴だけを使った場合、SVM が他の 4 つの学習アルゴリズム、すなわち Nearest Neighbor 法、ナイーブベイズ、決定木 および、Winnows の Sparse Network よりも、文の分類において高い性能を持つことが示されている。小林ら[3]はランダムフォレスト、語の頻度および、品詞タグ付された n-グラムをフォーマルな英語学術論文の質を判定する指標として使用した。彼らは、この方法を使用して、コーパス中の論品文の品質の良否を 77.75%の精度で判定している。

3. 原文修正文対のベクトル化

英文誤りの分類推定評価実験のための基礎データを以下の手順で構築した。まず、Lang8 にある英文日記から参加者による修正がある文をランダムに 500 個選んだ。ただし、これらの例には、文の大半を書き直したものもあった。そこで、そのようなものを取り除いた 399 件の例を対象とした。つまり、分析対象は文ではなく、誤りを含む原文とそれを修正した修正文の対を対象である。本稿では それらの対象について検索エンジン GETA[*1] を使って専用構築をした。具体的には、まず、原文と修正文に現れる単語を索引語として登録した。単語の誤りも分析 対象とするので、索引作りで通常行なるステミングは行わないこととした。

Lang8 では、原文に対する他の参加者による修正はというタ グとして記述されている。このタグのパラメーターxxx の部分には、単語の削除を表す sline の他に、f_bold,f_red,f_blue という表示の色を示す属性が使われている。しかし、全ての修正にタグで記述されている訳ではなく、また、統一的にタグ属性が使われている訳ではなかった。そこで、本稿ではこれらの修正タグを利用することは諦めた。その代わりに、原文と修正文にアラインメントを適用することで、修正部分を抽出した。 Lang-8 のデータは、例えば、下のような原文と修正文の対になっている。

表 1　原文と修正文の例

原文	I woke up alone, with lose memory, lying on the white beach, not knowing where I was.
修正文	I woke up alone, with no memory, lying on a white beach, not knowing where I was.

この例では、原文の「lose」ならびに「the」が、修正文ではそれぞれ「no」ならびに「a」 となっている。修正としては、「lose の削除(delete)」「the の削除(delete)」 「no の挿入(insert)」「a の挿入(insert)」の操作によりアラインメントできる。構築した検索エンジンでは、この例については、修正を表す d:no, d:the, i:no, i:a も一般の単語と同様に索引語として登録する。削除と挿入の区別なく修正(edit)とみなした索引語 e:no,e:the,e:no,e:a も登録した。

次に、英語を母国語とする第一著者が、それぞれの文に 42 種類の誤りのどれに 該当するか分類を行った。上の例については、誤りの種類としては 38 番の Article errors と、41 番の Nagation という 2 種類の誤りがあると判定した。これらの誤り分類(category)は、c:38 ならびに c:41 という索引語として登録した。この結果、先ほどの例は、エラー分類、修正単語、一般単語の３種類の索引語によりベクトル化され、i:a,d:the,e:a,e:the から Article についてのエラーということが推定できる。また、i:no,e:no から否定に関連するエラーも推定できる。このように 一般の単語だけでなく、修正情報を用いることで、エラー推定が可能になる と考えた。

*1 http://geta.ex.nii.ac.jp/geta.html

表 2　例文の索引語

c:38/ c:41
d:lose/ d:the i:no/ i:a
e:the /e:lose/ e:a /e:no
the/ a/ woke/ no/ not/ on/ white/ memory/ with/ lying/
beach/ up/ i/ knowing/ where/ alone/ was/ lose/

このような索引語によって専用検索エンジンを構築した。なお、誤り分類推定に は、c:38 や c:41 という分類情報は利用していない。

4. 英作文の誤り分類

前章で述べた通り、無作為に抽出した 500 英語添削例のうち、全文を書き直したものを除いた 399 例を誤りパターン分類のために使用した。

まず、これらの添削例を Kroll[6] および Weltig[7] に基づき 42 種類の誤りに手作業で分類した。 なお、[6]および[7]では、それぞれ細部が異なる分類を採用しているため、両者を統合した誤り分類表を使用した。

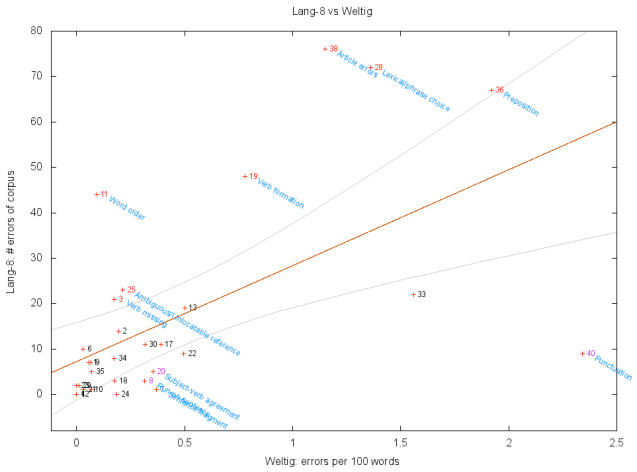

図 1　Lang-8 と Weltig の誤りの相関

従来の研究(Kroll [6] and Weltig [7])における各種誤りの発生頻度と、Lang-8 における誤り発生頻度について、線形回帰分析を行い相関を調べた。その結果、Lang-8 における誤り発生頻度は、Kroll の教室内、Kroll の自宅、および Weltig の調査における誤り発生頻度と有意な相関を示した（$p < 0.05$)。

表 3　誤り発生頻度の線形回帰分析（従来の研究と Lang-8)

	Kroll (教室)	Kroll (自宅)	Weltig
r^2	0.6351	0.6409	0.5834
t	4.3509	4.4179	3.8011
p	0.0002	0.0001	0.0007
y	2.9376 + 4.2918x	4.9722 + 3.6384x	7.2613 + 21.1171x

ネイティブスピーカーによる添削例では、しばしば、一つの文中に複数の種類の誤り訂正が含まれる。これを考慮に入れ、複数の種類の誤り修正を含んでいる添削例は複数の誤りパターンを持っているものとして分類した。いくつかの添削例では、語句の訂正についてコメントしたり、複数の訂正候補を示唆していた。これらは、「語句の選択」(Lexical/phrase choice)誤りとして分類した。

これらの相関関係を用いて、95%の信頼区間外にある誤り分類を識別した。つまり、従来の研究と Lang-8 の例とで発生頻度が異なる誤り分類を抽出した。3 つの 回帰分析において、合計 22 種類の誤りが 95%の信頼区間外にあった。3 つの分析すべてにわたって 95%の信頼区間外にプロットされた誤りは 11 種類であった。

これら 11 種類の誤りは、Lang-8 環境でのライティングにおける誤り発生および訂 正のパターンと、Kroll や Weltig の研究で使われたより伝統的な教育環境におけるパターンに違いがあることを示唆している。この違いは、学生のモチベーション、 ライティングの題材、学生の個人的特質（年齢、社会的経済的背景）等に起因 すると思われる。

表 4　Lang-8 において 95%信頼区間外にプロットされる誤り種類

	More freq. in Lang-8		Less freq. in Lang – 8
#	Error Cat.	#	Error Cat.
3	Verb missing	7	Sentence fragment
11	Word order	8	Run-on sentence
19	Verb formation	20	Subject-verb agreement
25	Ambiguous/Unlocatable reference	40	Punctuation
28	Lexical/phrase choice		
36	Preposition		
38	Article errors		

5. SVM によるエラー分類推定の性能評価

下の表 5 は 399 件全てのデータを学習データとして SVM を適用したときのエラー 分類推定性能である。なお、この表 5 では F 値でソートしている。36(Preposition), 42(Spelling), 2(Subject formation), 28(Lexical/phrase choice)では 9 割以上の性能で推定できている。ただし、これは全てのデータを使っているので汎化性能の評価ではない。

表 5　全データを利用したエラー分類推定性能

分類	prec	recall	F	accuracy
36	0.9310	0.9643	0.9474	0.9850

42	0.9773	0.8958	0.9348	0.9850
2	1.0000	0.8571	0.9231	0.9950
28	0.8696	0.9677	0.9160	0.9724
38	0.2698	1.0000	0.4250	0.5388
19	0.1845	1.0000	0.3116	0.5238
11	0.1201	1.0000	0.2145	0.3208
33	0.0955	1.0000	0.1743	0.5013
25	0.0806	1.0000	0.1493	0.4286
3	0.0599	1.0000	0.1131	0.2531
17	0.0521	1.0000	0.0990	0.5439
13	0.0492	1.0000	0.0939	0.3709
6	0.0488	1.0000	0.0930	0.5113
37	0.0478	1.0000	0.0913	0.5013
30	0.0461	1.0000	0.0881	0.4812

表 6、図 2、図 3 は 399 件のデータをランダムに 10 個に等分割し、9 割で学習し残りの 1 割でのテストを 10 回行いその平均を求めた 10 分割交差検定の結果である。 表 6 では 1 列目にデータの個数も合せて表示している。いずれのエラー分類でも、F 値は 4 割にも満たない。

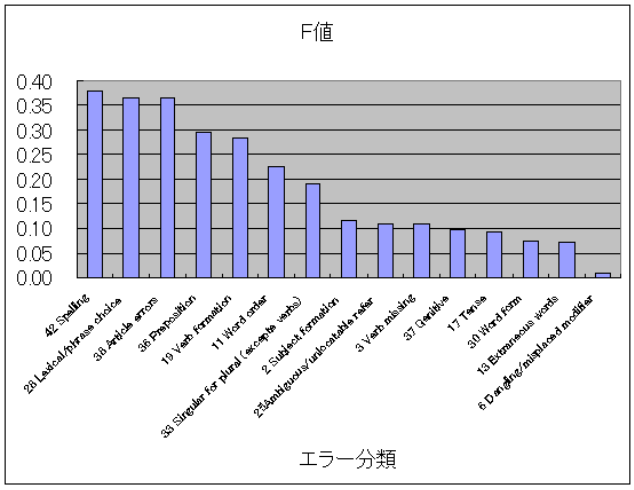

図 2　エラー分類ごとの推定性能（F 値、10 分割交差検定）

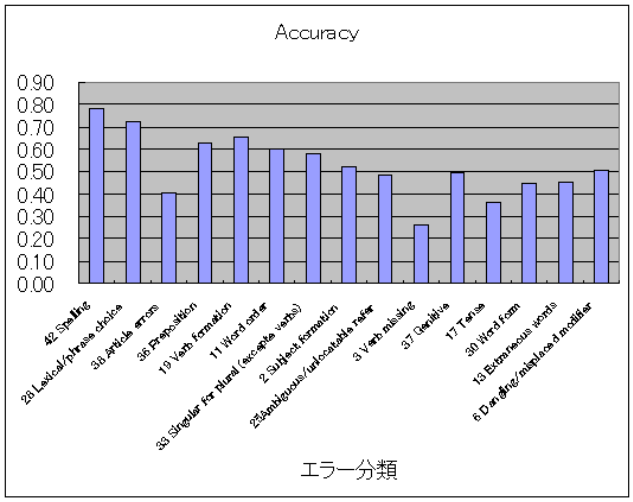

図 3　エラー分類ごとの推定性能
（Accuracy、10 分割交差検定）

データ数と F 値、ならびに Accuracy の相関をプロットした図 4、図 5 では、サンプル数が多いエラーの種類については性能も高くなるという正の相関がみられる。図 4 からは、夫々のエラーの種類について人手による識別事例を 100 件程度準備すれば、8 割の F 値が期待できると予想できる。

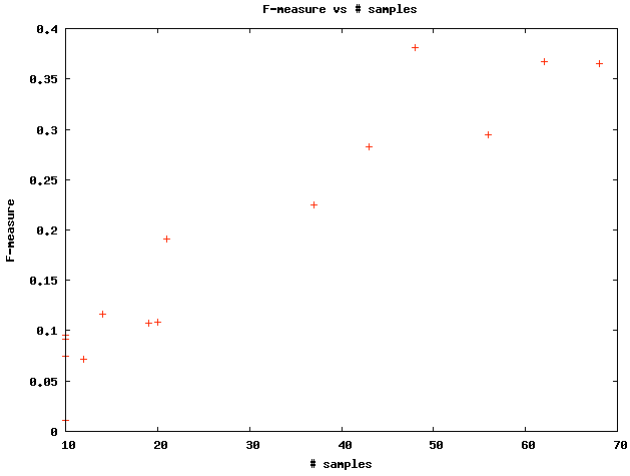

図 4　F 値とデータ数の相関

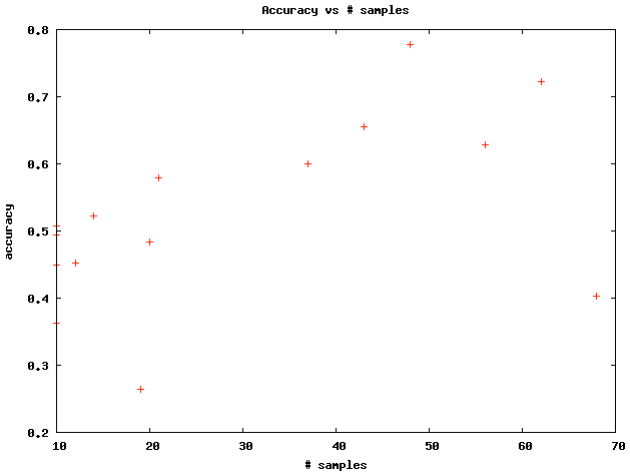

図 5　Accuracy とデータ数の相関

表 6　10 分割交差検定によるエラー分類推定性能

	error type	#sum	prec	recall	F	accuracy
42	Spelling	48	0.4153	0.3906	0.3807	0.7780
28	Lexical/phrase choice	62	0.3109	0.5206	0.3672	0.7218
38	Article errors	68	0.2265	0.9857	0.3652	0.4023
36	Preposition	56	0.2049	0.5742	0.2948	0.6288
19	Verb formation	43	0.1865	0.6881	0.2828	0.6547
11	Word order	37	0.1472	0.6514	0.2248	0.5999
33	Singular for plural	21	0.1129	0.8000	0.1910	0.5796
2	Subject formation	14	0.0758	0.3333	0.1169	0.5217
25	Ambiguous/unlocatable refer	20	0.0687	0.2833	0.1087	0.4843
3	Verb missing	19	0.0585	0.8250	0.1077	0.2647
37	Genitive	10	0.0539	0.4667	0.0957	0.4941
17	Tense	10	0.0588	0.4167	0.0917	0.3633
30	Word form	10	0.0418	0.3833	0.0750	0.4491
13	Extraneous words	12	0.0385	0.6500	0.0718	0.4516
6	Dangling/misplaced modifier	10	0.0063	0.0333	0.0105	0.5078

6. エラー分類のための特徴語

全てのデータについて SVM を適用して得られるモデルから、単語とタグのスコアを求めることができる。38(Article errors)では、単語そのもより、タグが付いた e:the,i:the,e:a,i:a などが特徴語となっている。36(Preposition)でも、タグが付いた i:in,e:in,d:at,e:for,e:at,e:on,i:on などが特徴語となっており、修正情報がエラー識別に有効になっていることが確認できる。19(Verb formation)の特徴語からは、"ing"が特徴的な誤りとなっていること が予想できる。42(Spelling) の "e","e:e","i:e" については、 conv-a-rsation, ev[e]ryone,などの間違いであった。

表 7　SVM を適用して得られるモデルから、単語とタグ

err	特徴語	
42	Spelling	shopping e went e:e i:e phrase china day friend what
28	Lexical/phrase choice	which m it am would student in d:in here girl
38	Article errors	e:the i:the e:a the i:a a man e:A university e:This
36	Preposition	i:in e:in d:at at e:for e:at e:on on i:on two
19	Verb formation	i:ing e:ing ing didn e:to entrance d e:eat d:eating collage

7. まとめと今後の課題

本稿では、Lang-8 に書かれた英文日記から、英語を母国語とする第一著者が人手で(Kroll[6], Weltig[7])による作文誤りのどれに該当するか識別を行った。原文と添削文のアラインメントにより、削除、挿入された単語を抽出し、修正タグを合わせて索引語とした。 原文、修正文に含まれている単語だけでなく、これらの修正情報も属性として利用した。こうして得られた学習データに対しパターン分類器 SVM を適用して、誤り分類の推定を行い、その推定性能を評価した。具体的には 399 件のデータをランダムに 10 個に等分割し、9 割で学習し残りの 1 割でのテスト 10 回行いその平均を求めた 10 分割交差検定を行った。いずれのエラー分類でも、F 値は 4 割にも満たない。しかし、準備したデータ数と F 値、ならびに Accuracy の相関をプロットした結果、サンプル数が多いエラーの種類については分類推定性能も高くなるという正の相関がみられた。夫々のエラーの種類ごとに、人手による識別事例を 100 件程度準備すれば、8 割の F 値が期待できると予想できる。そこで、学習データを増やすことで、性能向上を目指す予定である。しかし、与えられた文の誤りを 42 種類のどの誤りに該当するか判定する作業は容易ではない。そこで、性能は 4 割以下だが、本稿で得られた判別器を利用することで、予想エラー分類のランキングとして表示すれば、人手による判定の効率が上がると予想している。

参考文献

1) 鈴木潤, 佐々木裕, 前田英作: 単語属性 N-gram と統計的機械学習による質問タイプ同定, 情報処理学会論文誌, 44(11), 2839-2853 (2003).
2) 平野孝佳, 平手勇宇, 山名早人: 検索エンジンを用いた英文冠詞誤りの検出, DBSJ Letters, 6(3) (2006).
3) 小林雄一郎, 田中省作, 冨浦洋一: N-gram を素性とするパターン認識を用いた英語科学論文の質判定, 研究報告情報基礎とアクセス技術(IFAT), 12(1) (2012).
4) 谷本太郁由, 太田学: 検索エンジンの検索結果数に基づく英文誤り検出に関する検討, DEIM Forum 2012, 9(1) (2012).

5) Karatzoglou, A., Meyer, D., Hornik, K.: Support vector machines in R. Journal of Statistical Software, 15 (9), pp. 1-28 (2006).
6) Kroll, B.: What does time buy? ESL student performance on home versus class compositions, In B. Kroll (Ed.), Second language writing: Re- search insights for the classroom (pp. 140-154), Cambridge: Cambridge University Press (1990).
7) Weltig, M. S.: Effects of language errors and importance attributed to language on language and rhetorical-level essay scoring, Spaan Fellow Working Papers in Second or Foreign Language Assessment Volume 2 2004,1001, 53 (2004).
8) Yin, C., Hirokawa, S., Flanagan, B., Suzuki, T., Tabata, Y.: Mistake Discovery and Generation of Exercises Automaticity in Context, Proc. of LTLE2012 (2012).
9) Zhang, D., Lee, W. S.: Question classification using support vector machines, In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 26-32) (2003).