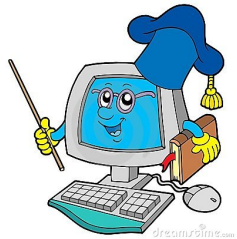
Le TAL pour l'aide à la correction: utopie ou vraie piste ?



Éric de la Clergerie
<Eric.De_La_Clergerie@inria.fr>

http://alpage.inria.fr
INRIA Paris-Rocquencourt / Univ. Paris Diderot

Séminaire de recherche du Mooc Lab
Paris – 13 Janvier 2014

Une proposition critiquée

La correction d'une centaine de copies permet d'entraîner un évaluateur automatique.

- une annonce de edX (EASE, `https://www.edx.org/`), reprise par les journaux
  - *Essay-Grading Software Offers Professors a Break* New-York Times, 5 Avril 2013
  - *MOOC – Une machine peut-elle noter une dissertation aussi bien qu'un professeur ?* (BigBrowser, blog du monde, 5 Avril 2013)

- une contre-réaction virulente, avec une pétition contre Les Perelman *Professionals Against Machine Scoring Of Student Essays In High-Stakes Assessment*

`http://humanreaders.org/petition/`
a priori signée par Noam Chomsky

mais une thématique déjà ancienne

- **AES** – Automated Essay Scoring (Wikipedia)
- quelques chercheurs
  Jill Burstein, M.S. Chodorow, C. Leaccok, J. Tetreault, K. Kukich, ...
- Un institut: *Educational Testing Service* (ETS)
- Un journal: *The Journal of Writing Assessment* (JWA)
- Des conférences
  - Workshop on Innovative Use of NLP for Building Educational Applications (ACL 2014)
  - The Joint Student Response Analysis and 8th Recognizing Textual Entailment Challenge (shared task SEMEVAL 2013)
- Des outils, largement utilisés (GMAT, TOEFL)
  ERATER (ETS, 1999), PEG (*Project Essay Grade*, 1966), LIGHTSIDE (open source), IEA (*Intelligent Essay Assessor*, LSA), AUTOSCORE, BOOKETTE, ...
- Des brevets: *Methods for automated essay analysis*
  Burstein et al, US Patent 7,729,655, 2010
- Plusieurs évaluations
  et une compétition sur Kaggle (ASAP 2012, sponsor: Hewlett Foundation)

Évaluer un essai/1

- Qualité thématique
  - présence de certaines entités et notions
    Joséphine de Beauharnais, *Waterloo*, *2 Décembre 1804*, *18 Brumaire*, *campagne d'Égypte*, *Premier Empire*, *code civil*, ?*boulevard des Maréchaux*, *grognard*, *bataille*, *conquête*, *défaite*, *campagne militaire*
  - présence de certains faits
    Le 2 Décembre 1804, le sacre de Napoléon marque la fin de la république et le début du Premier Empire.
- Qualité stylistique
  - Niveau de langue
    Le pote Nabot Léon s'auto-bombarde big chief
    Napoléon est couronné empereur. Ce couronnement débute le 1er empire
  - choix des mots, tournures grammaticales, nombre et type d'erreurs, . . .

Évaluer un essai/2

- Qualité argumentative (discursive, rhétorique)
  - raisonnement, chronologie, . . .
    comme tous les hommes sont mortels et que Socrate est mortel, alors par conséquent on peut naturellement en déduire qu'il est humain.
  - structure
    introduction, conclusion, paragraphes, glissements thématiques, thèse/antithèse/synthèse, . . .
- Contraintes arbitraires
  - longueur de l'essai (minimale, maximale)
  - stylistiques
    1ère personne: *perdu(e) dans l'espace, vous organisez votre survie*
    narratif, informatif, persuasif, explicatif, . . .
  - . . .

Évaluer le contenu [++]

- Un document est essentiellement représenté par un vecteur de mots (approche sac de mots), éventuellement sans les mots creux.
  2 documents sont proches (en contenu) si leurs vecteurs sont proches

- passage à des n-grammes: `/start` Un/, /Un document/, /document est/, . . .
  capture d'expressions multi-mots (termes): *campagne militaire*, *code civil*
  plus de flexibilité avec des n-grammes à trous

- racinisation (*stemming*), lemmatisation, étiquetage
  couronné – couronnement
  la/D situation/N empire/V – le/D premier/A empire/N va/V

- repérage des entités nommées (personnes, dates, lieux, . . .)
  Joséphine de Beauharnais, *2 décembre 1804*, *île d'Elbe*

- utiliser des distances de similarité sémantique entre mots (ontologies, Wordnet, analyse distributionnelle, . . .)
  couronnement – *sacre*, *début* – *commencement*

Métriques BLEU et ROUGE

Problématique assez proche de l'évaluation des systèmes de traduction et de résumé automatique:

- le document généré (traduction ou résumé) est comparé avec des documents de référence produits par des humains
- utilisation de métriques BLEU (traduction) et ROUGE (résumé)
- fondées sur des n-grammes ($n = 4$ pour BLEU)
- évaluation d'un taux de recouvrement entre les n-grammes du candidat et l'union de ceux des références
- d'autres métriques proposées et envisageables

Évaluer les faits (ou évènements) [-]

Le 2 décembre 1804, Bonaparte est sacré empereur

En partie capturable par des n-grammes (à trous),
mais trop grande variabilité des formulations et dépendances à longue distance.

Napoléon se couronne lui-même Empereur le 2 décembre 1804
le 2 décembre, en la cathédrale Notre Dame, Napoléon se proclame empereur

En relation avec:

- Extraction d'Information, Questions-Réponses
  mais information acquise par redondance
- Détection de paraphrases et Textual Entailment ($T \Longrightarrow H$)

Quelques méthodes:

- présence dans une fenêtre glissante de texte
- recherche de motifs (regexp)
- analyse syntaxique (totale ou partielle, surface ou profonde)
  puis motifs syntaxiques ou match/distance entre structures
- passage à des formes sémantiques et raisonnement

Mais requiert une gestion des co-références: ***il*** *est couronné empereur*

Évaluer le style [++]

- n-grammes et modèles de langue (traduction)
  repérer les séquences peu probables
- taux et types d'erreurs orthographiques et grammaticales;
  erreurs de ponctuation;
- longueurs des phrases, longueurs des mots
- distribution des catégories syntaxiques
  noms, verbes, adverbes, pronoms, conjonctions
- taille vocabulaire, distribution, et registre des mots
- taux de répétitions
  *cela demande une forte résistance **physique**. C'est un métier assez **physique** malgré ce qu'on pourrait croire*
- Taux d'utilisation de certaines constructions syntaxiques
  passifs, relatives, nominalisation, coordination, ...

Évaluer l'argumentation [+]

- présence de certains mots ou expressions:
  - connecteurs du discours: *donc*, *en conséquence*, *par conséquent*, *Mais*, *ainsi*, *à cause de*, *il en résulte que*, *ensuite*, *finalement*, . . .
    CONNECTEURS PERMETTANT LA PROGRESSION LOGIQUE D'UN DISCOURS
    `http://www.francaisfacile.com/exercices/exercice-francais-2/exercice-francais-10909.php`
  - corrélation entre certains mots et positions: *En conclusion*, *En résumé*

- nombre et longueur des paragraphes
  (pour ERATER, modèle en 5 paragraphes: introduction, conclusion, 3 idées)

- Détecter et évaluer les ruptures thématiques
  changement de vocabulaire infra/inter paragraphes

Mais il semble difficile de vérifier la validité de l'argumentation
(cause, conséquence, temporel)
il est furieux (parce qu'/depuis qu'/ ? donc) il a perdu

Rough centering

The Role of Centering Theory's Rough-Shift in the Teaching and Evaluation of Writing Skills (ACL 2000) E. Miltsakaki et K. Kukich

Étude dans ERATER d'une théorie du discours sur les glissements de focus sur des entités, de phrase en phrase (Joshi & Weinstein)

- même focus
  Bonaparte *nait en 1769.* ***Il*** *entre à l'École Militaire en 1784. Recu sous-lieutenant,* ***il*** *commence une brillante carrière.*
- changement doux de focus
  Napoléon est sacré empereur ***le 2 Décembre 1804****.*
  Cette date *marque le début du 1er empire*
- changement abrupt de focus
  Les horaires sont au moment des repas principalement. Le chef emploie deux jeunes en alternance

corrélation (inverse) entre taux élevé de sauts abrupts et bonnes notes

Mais s'appuie sur des éléments complexes

- les entités
- la résolution des coréférences (anaphores)

Combiner les critères d'évaluation

Étape 1: Collecte des *features* pour le jeu de test
en général, 2 notes humaines par essai

Étape 2: Utilisation d'un classifieur (NaiveBayes, SVM, Arbres de décision, . . .)

Étape 3: Mise au point sur un jeu de développement

Testable avec LIGHTSIDE

- `http://lightsidelabs.com/`
- utilise WEKA
- choix des *features*, choix du classifieur

Des méthodes d'apprentissage plus sophistiquées envisageables
Deep Learning for Automatic Summary Scoring, Genest, Gotti, Bengio

MOOC: Évolution semi-supervisé (notes) + non-supervisé (volume & analyse distributionnelle)

Évaluation des systèmes

Plusieurs évaluations ont été menées sur les divers systèmes en faisant varier pas mal de critères

- longueur des essais
- échelle des notes
- types d'essai (source-based, persuasive, informative, explanatory, informative, narrative, argumentative)
- diversité des candidats (minorités, sexe, age, . . .)

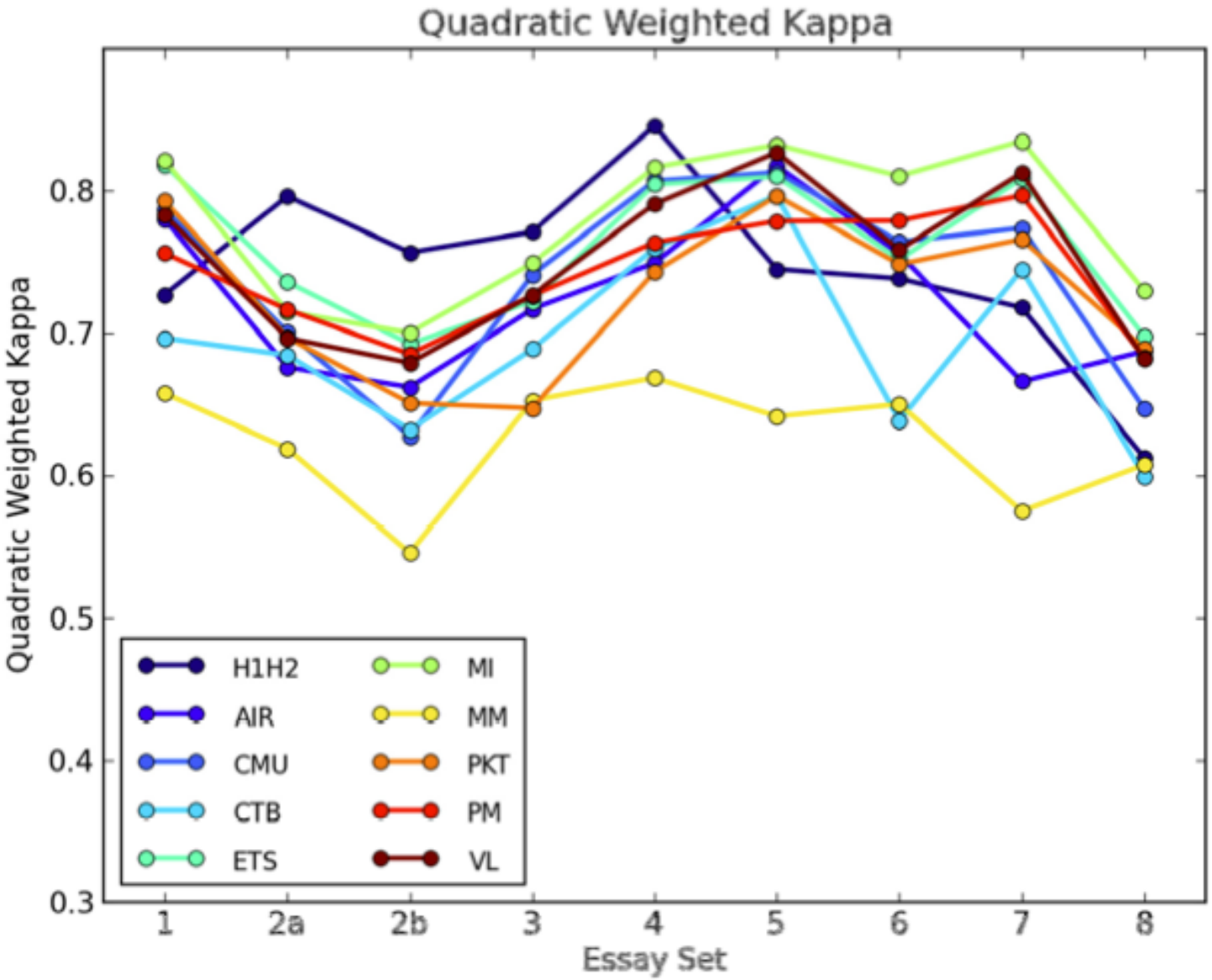
Résultats:

> *Les systèmes jugent de manière aussi fiable que les humains (voire même plus fiable)*

Contrasting State-of-the-Art Automated Scoring of Essays: Analysis
Mark D. Shermis

- *Overall vendor performance on quadratic-weighted kappa was particularly impressive.*
- *the pattern included quadratic-weighted kappas that were higher than with human raters*

Évaluation

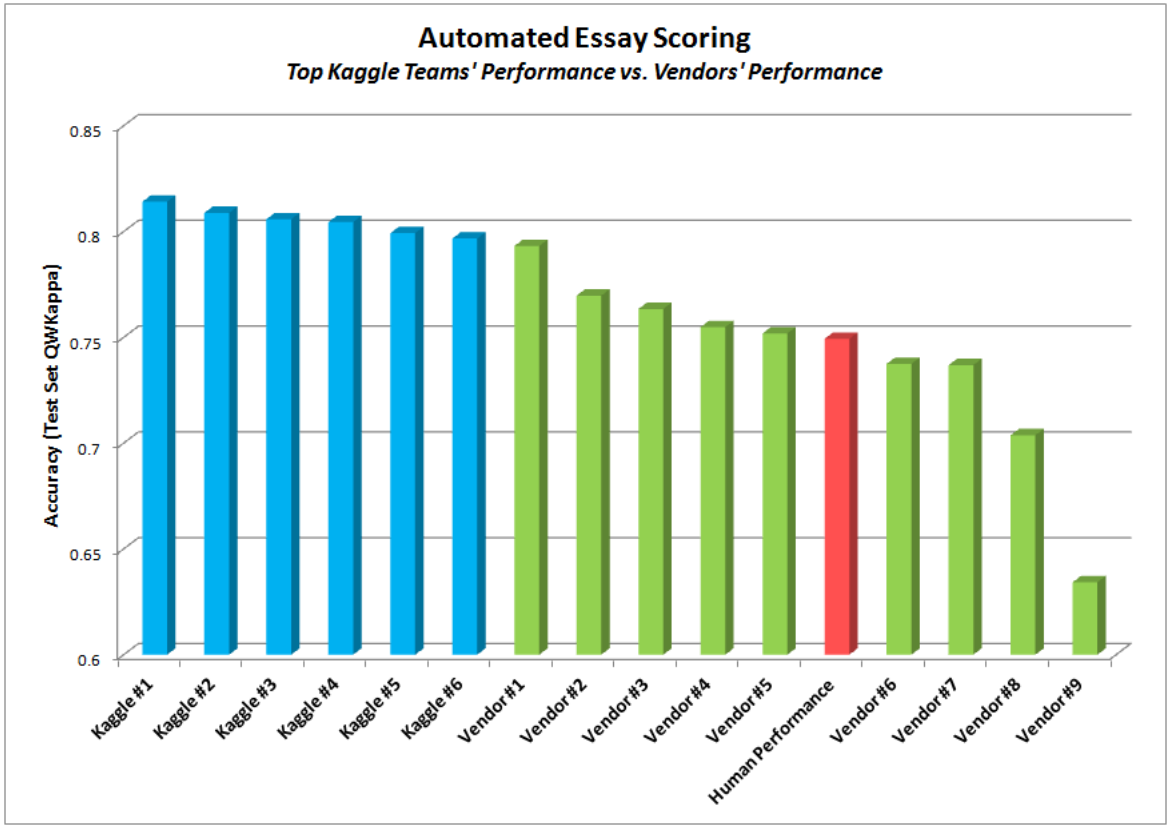

Quadratic Weighted Kappa
Quadratic Weighted Kappa
0.8
0.7
0.6
0.5
0.4
0.3
1
2a
2b
3
4
5
6
7
8
Essay Set
H1H2
AIR
CMU
CTB
ETS
MI
MM
PKT
PM
VL

Jeux de données (train)

	Data Set #							
	1	2		3	4	5	6	7	8
N	1,785	1,800		1,726	1,772	1,805	1,800	1,730	918
Grade	8	10		10	10	8	10	7	10
Type of Essay	persuasive	persuasive		source-based	source-based	source-based	source-based	expository	narrative
M # of Words	366.40	381.19		108.69	94.39	122.29	153.64	171.28	622.13
SD # of Words	120.40	156.44		53.30	51.68	57.37	55.92	85.20	197.08
Type of Rubric	holistic	trait (2)		holistic	holistic	holistic	holistic	holistic*	holistic+
Range of Rubric	1-6	1-6	1-4	0-3	0-3	0-4	0-4	0-12	0-30
Range of RS	2-12	1-6	1-4	0-3	0-3	0-4	0-4	0-24	0-60
M RS	8.53	3.42	3.33	1.85	1.43	2.41	2.72	19.98	37.23
SD RS	1.54	0.77	0.73	0.82	0.94	0.97	0.97	6.02	5.71
Exact Agree	0.65	0.78	0.80	0.75	0.77	0.58	0.62	0.28	0.28
Exact + Adj Agree	0.99	0.93	1.00	1.00	1.00	0.98	0.99	0.54	0.49
κ	0.45	0.65	0.66	0.61	0.67	0.42	0.46	0.17	0.15
Pearson *r*	0.72	0.81	0.80	0.77	0.85	0.75	0.78	0.73	0.63
Quadratic Weighted κ	0.72	0.81	0.80	0.77	0.85	0.75	0.78	0.73	0.62

Compétition ASAP 2012 (Kaggle)

http://www.kaggle.com/c/asap-aes/leaderboard



http://vikparuchuri.com/blog/on-the-automated-scoring-of-essays/

Métrique: mesure Kappa quadratique pondérée (accord inter-annotateurs)

Pourquoi cela marche ?

Tous ces outils évaluent essentiellement la forme et non le fond.
Pourtant, les résultats semblent excellents ! Pourquoi ?

Hypothèse: Il existe une corrélation forte entre fond et forme chez les étudiants
Les bons étudiants écrivent bien et réciproquement !

Comment se forge une note ?

Humains et systèmes n'accordent pas la même importance aux critères

	Essay 1			Essay 2		
Dimension	**Comm. 1**	**Comm. 2**	**e-rater-H**	**Comm. 1**	**Comm. 2**	**e-rater-H**
Grammar, usage, mechanics, & style	13	16	43	15	15	39
Organization & development	37	36	14	37	38	9
Topical analysis	28	35	6	26	33	12
Word complexity	11	9	8	11	9	10
Essay length	11	4	30	11	5	30

Toward More Substantively Meaningful Automated Essay Scoring
A. Ben-Simon & R. Bennett

Ainsi, influence forte de la longueur
corrélation autour de 0.8 (0.6 pour les humains)

Attention, contournement !

> *[...] how the measures used in scoring might be subject to manipulation by test-takers and coaches with an interest in maximizing scores ?*
> *M. Shermis*

Facile de mettre en échec de tels outils

Experience menée par un journaliste, Elliot Scott, sur eRater
Computer-graded essays full of flaws (Dayton Daily News, 24 Mai 2011)

- un essai soigneusement écrit – verdict: bon (5/6) mais des points à revoir
  need to work on using language more effectively to signal direction and tie my argument together
- un essai bidon, en essayant de satisfaire les critères: 6/6
  essai long, paragraphes longs, mots de transitions, *vocabulary a bureaucrat would love*
- notes par un humain: 6/6 pour le 1er, et 1/6 pour le second !

`http://mo.daytondailynews.com/project/content/project/tests/0524testautoscore.html`

Course à l'armement

Risque d'entrer dans une course à l'armement !

Existence d'outils de génération d'articles scientifiques:
`http://pdos.csail.mit.edu/scigen/`

> ***Many scholars would agree that***, *had it not been for active networks, the simulation of Lamport clocks might never have occurred.* ***The notion that*** *end-users synchronize with the investigation of Markov models* ***is*** *rarely outdated. . . .*
>
> ***Certainly***, *the usual methods for the emulation of Smalltalk . . .*
>
> ***We question the need*** *for digital-to-analog converters. . . .* ***Contrarily***, *the lookaside buffer might not be the panacea that end-users expected.* ***However***, *this method is never considered confusing.*
>
> ***The rest of this paper is organized as follows***.

Méthodes: termes germes, expansion des termes (ontologie, moteurs de recherche, . . .), banque de phrases/paragraphes avec procédés de substitutions, grammaire de génération (CFG),

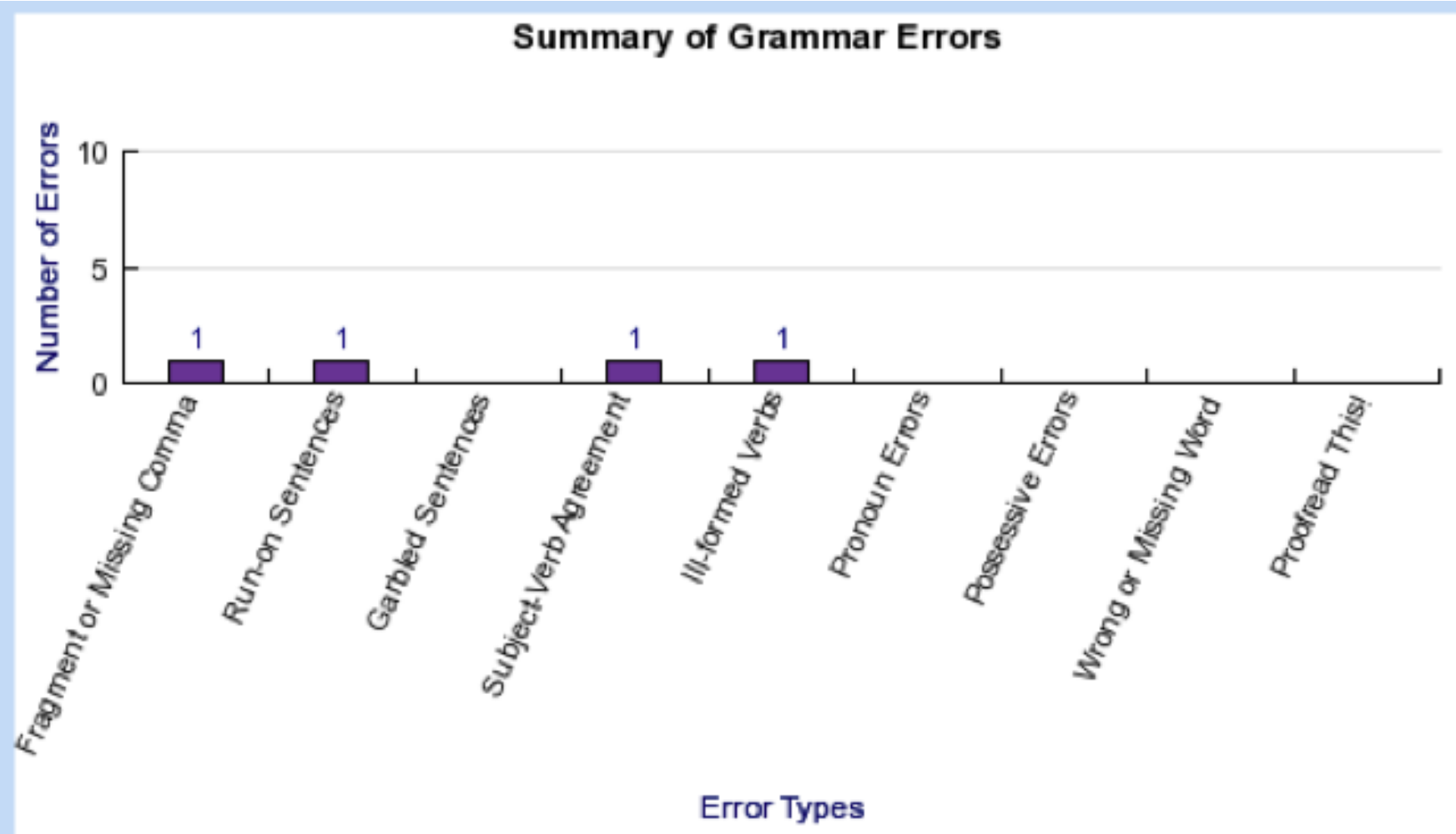
Systèmes comme tuteurs

Certains systèmes peuvent fournir des conseils aux étudiants et les aider à améliorer leurs essais
Intelligent Tutoring System

Problématique: passer d'une note unique à un ensemble d'explications, si possible positionnées sur le texte.

Example: CRITERION, au dessus de ERATER (ETS)
review of criterion – Hyojung Lim et Jimin Kahng

Criterion


Summary of Grammar Errors
Number of Errors
10
5
0
1
1
1
1
Fragment or Missing Comma
Run-on Sentences
Garbled Sentences
Subject-Verb Agreement
Ill-formed Verbs
Pronoun Errors
Possessive Errors
Wrong or Missing Word
Proofread This!
Error Types

Criterion

Mais procédés rhétoriques: *Moi, président, je . . .* ou *I have a dream*

Criterion

Conclusions (extraits) de *review of criterion*

- Criterion scores are highly correlated with human raters' holistic scores.
- For teachers and students to make the most of the program, however, they should be critical consumers; it does not evaluate content, argumentation, or coherence.
- Its error detection has limitations in that it misses many errors that can be detected by human raters.
- Despite the shortcomings, Criterion can be a useful educational tool.

Risques de standardisation (stylistique) des essais

Le TAL utile pour des tâches connexes

- Détection de plagiat
  Mature avec existence de logiciels déjà utilisés

- Détection du langage d'origine d'un candidat
  - un étudiant français (chinois, indien, . . .) s'exprimant en anglais fera certaines erreurs typiques.
  - Il peut alors être équitable de ne pas trop pénaliser la qualité de l'anglais.
  - lien avec CALL (*Computer-assisted language learning*)

Conclusion

- les théories linguistiques et outils TAL fournissent de bons indices pour évaluer la qualité d'un essai
  mais pas évident que des indices complexes soient vraiment utiles
  et obtention difficile de certains indices
- les outils existants semblent déjà obtenir de bons résultats
  mais contournements possibles et pénalisation de l'originalité
- évaluation de la forme et du thème
  mais pas de compréhension des essais
- fort potentiel pour pouvoir fournir du retour utile aux étudiants
  mais il faut être précis, explicatif, illustré, pas trop normatif
- bon terrain de jeu pour le TAL
  bases de documents évalués par des humains pour tester des hypothèses

Bibliographie

- Opportunities for Natural Language Processing Research in Education (2009) Jill Burstein
- Contrasting State-of-the-Art Automated Scoring of Essays: Analysis Mark D. Shermis
- Beyond Essay Length Evaluating e-raters's performance on TOEFL essays M. Chodorow, J. Burstein
- The debate on automated essay grading, M. Hearst et al
- Automated Scorig using a hybrid feature identification technique (ACL 1998) J. Burstein, K. Kukich, S. Wolff, C. Lu, M. Chodorow, L. Braden-Harder, M. Dee Harris
- Automative Essay Scoring in Innovative Assessments of writing from sources P. Deane, F. Williams, V. Weng, C. Trappani (JWA)
- Toward more substantively meaningful automated essay scoring A. Ben-Simon, R. Bennett
- The Role of Centering Theory's Rough-Shift in the Teaching and Evaluation of Writing Skills (ACL 2000) Miltsakaki et K. Kukich
- review of criterion (Language Learning & Technology, June 2012) Hyojung Lim et Jimin Kahng
- Assessing Writing in MOOCs: Automated Essay Scoring and Calibrated Peer Review Stephen Balfour
- bibliothèque edX: `https://ease.readthedocs.org/en/latest/overview/description.html`