From Linguistic Properties to Extra-Linguistic Properties

Hans van Halteren
Radboud Universiteit Nijmegen

hvh@let.ru.nl

Part 1: Plagiarism by students

Teachers sometimes have to judge a student on the basis of something written outside a controlled environment.

The student may have decided to "reuse" existing material: ***PLAGIARISM***

Let's look at it from the student's point of view:

De situatie

Probleem! Je moet je werkstuk voor morgen eigenlijk nog afmaken. Maar je hebt met je vrienden afgesproken om over een half uur te gaan stappen.

Oplossing? Je zoekt iets wat voldoende past bij de opdracht. Mogelijke bronnen: andere leerlingen en Google. Dat moet nog wel lukken in een half uur.

Richtlijnen voor je keuze

Maar van wie moet je kopiëren?
Optie 1, een andere leerling

Kwaliteit
- De beste die je durft te vragen
- Een niet zo goede maar wel creatieve
- Een middelmatige
- Een slechte, want dat is minder verdacht

Tijd
- Uit jouw groep
- Uit een andere groep
- Uit een ander jaar

Richtlijnen voor je keuze

De antwoorden:
een middelmatige, in ieder geval uit een andere groep en liefst uit een ander jaar

Waarom:
het hoofdpunt is ***niet opvallen***

Wanneer startte ik zelf controles: toen ik iets heel onwaarschijnlijks twee keer zag

Idem voor internet:
kies iets wat van jou zou kunnen zijn
("dit Engels is veel te goed voor onze studenten")

Richtlijnen tegen detectie

Is een goede keuze voldoende om aan detectie te ontsnappen:

- JA
- NEE

Richtlijnen tegen detectie

Is een goede keuze voldoende om aan detectie te ontsnappen:

- ***NEE***

Er is software voor
- vergelijken met andere leerlingen
- op internet zoeken naar bron

Dus hoe kunnen we de tekst aanpassen zodat we niet gesnapt worden?

Richtlijnen tegen detectie

Hoe kunnen we de tekst aanpassen zodat we niet gesnapt worden?

- Ideeen verzamelen (en op bord zetten)

En dan weer terug naar de tegenpartij:

Case 1: Limited sources

- specific assignment
- unlikely to have been done before

Only possible source: fellow students

Need to check: Overlap

Case 1: Limited sources

RUN: subscribe to service (Ephorus)

Also software available on internet, e.g
- Wcopyfind (http://www.plagiarism.phys.virginia.edu)

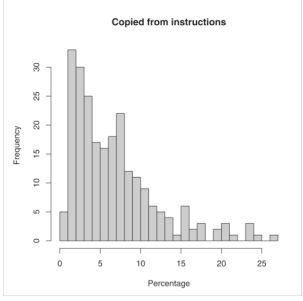
Overlap test, with parameters such as
- which percentage overlap should be reported
- how long/short can matching phrases be
- how many imperfections can be inside them
- what to do with punctuation/case/numbers/...
- even use of a word map!

Case 1: Limited sources

Problem: threshold setting

Special problem: students copy bits from task description

One-on-one overlap test not enough...



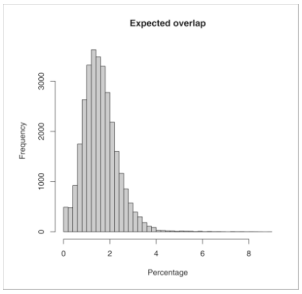
Case 1: Limited sources

My solution: Trigram overlap

- convert to ASCII and tokenize
- collect all trigrams present in text
- remove all trigrams from task description
- calculate percentage reused / all

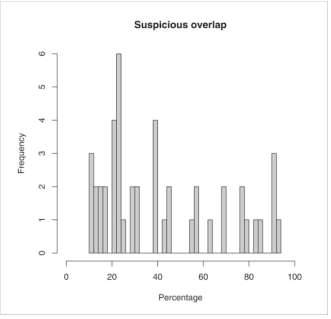
Case 1: Limited sources

Overlap for independent case studies



Case 1: Limited sources

Higher than normal overlap



Case 1: Limited sources

Threshold: 10%

Robust against:
- local rephrasings
- spelling error introduction
- reordering
- copying text from multiple sources

Making enough changes to fool the system is more work than doing the assignment.

Case 2: Source = Internet

- predictable or free assignment
- possibly done before/elsewhere
- at least possibility of unreported quotes

Possible source: anything on the internet

Need to check: Presence of foreign material

Case 2: Source = Internet

Obvious solution:

google for text fragments

Searching for a few 4- or 5-grams will suffice

(See also:
www.fdewb.unimaas.nl/eleum/plagiarism/plagiarism.htm)

Consequences for the student:

Verhulling: spelfouten

Spelfouten toevoegen helpt niet.

- werkt alleen plaatselijk: je zou een op de drie of vier woorden moeten aanpassen
- verzonnen spelfouten zijn vaak onnatuurlijk, en daardoor juist weer verdacht

Verhulling: tekst omgooien

- Zinnen of grotere stukken tekst verwisselen helpt niet, want trigrammen binnen de zinnen blijven gelijk
- Synoniemen uitwisselen helpt een beetje, maar je moet er weer heel veel
- Parafraseren helpt goed, als je maar genoeg verandert

Verhulling: vertalen

[komisch intermezzo]

Ideale manier van tekst omgooien: vertalen.
En kan automatisch!
(genoeg om te browsen, maar kwaliteit kan beter; b.v. Babelfish / Systran)

We vertalen de voorbeeldtekst
- van Nederlands naar Engels
- van Engels naar Duits
- van Duits naar Frans

Verhulling: vertalen

Resultaat: 6.54% overlap!

HOERA!
Detectie omzeild!

Tenminste?

Verhulling: vertalen

Resultaat: 6.54% overlap!

Die slaven communiceerden onderling in een mengelmoesje van woorden uit allerlei talen, vooral die van hun bazen.
IS GEWORDEN
Deze slaven waren wederzijds woorden van alle soorten talen mengelmoesje, in het bijzonder in de betrekking die van hun hoofden.

Helaas
- niet meer echt dezelfde inhoud
- niet meer echt Nederlands

(NB voor die twee bestaat trouwens ook software)

Verhulling: obscure bronnen

Als je dus iets wilt kopiëren zonder gesnapt te worden zou je het helemaal om moeten schrijven. Kun je het net zo goed zelf doen.

MAAR detectie (en bewijs) werkt alleen als de bron te vinden is

DUS gebruik een bron die niet
ófal eens ingeleverd is
ófop internet staat

Case 2: Source = Internet

Obvious solution:
google for text fragments
Searching for a few 4- or 5-grams will suffice

(See also:
www.fdewb.unimaas.nl/eleum/plagiarism/plagiarism.htm)

Problems:
- must check large number of text fragments
- source might not be accessible this way

In general: Source = Anything

If source unknown:

Need to check: Presence of foreign material

- don't try to find source
- try to determine if this student wrote this

Authorship Verification

Authorship Verification

Existing test:
"this English is much too good;
it cannot be produced by one of our students!"

- teacher can spot
- automated check: COLING2004 paper
  (show if there's time left)
- works mainly for foreign language

Part 2: Authorship Verification

My solution: Linguistic Profiling

Cf. ACL-2004 paper

Linguistic Profiling for Author Recognition and Verification

Hans van Halteren
Radboud University Nijmegen

hvh@let.ru.nl

The Task: General

Determine information about a text
on the basis of linguistic properties of the text

e.g.
- which genre / text type
- identity / age / gender of author
- classification for document routing
- level of certainty for information extraction

The Task: Approaches

Find properties you know are distinguishing
- use of function words
- frequency / presence of content words

But human insight may fall short, so

Linguistic Profiling
- Use all features you can think of
  (and are manageable)
- Let the system figure out which are useful
  for the task at hand

The Task: Specific

Example application area: Student essays

Is each written by the marked author?
- *Author Verification*

Can we assign author to unmarked essays?
- *Author Identification*
  (humanities: Authorship Attribution)
- Possibly *Author Sorting*
  (one student – one essay)

The Task: Evaluation

Experimentation is only useful if the results can be evaluated objectively!

Necessary:

- Material for which truth is known
- Measures which are appropriate for task

The Task: Measures

Basic measures
- False Accept Rate (FAR)
- False Reject Rate (FRR)

- Depend on threshold: FAR down = FRR up

But what do we want to optimize?

The Task: Measures

Basic measures
- False Accept Rate (FAR)
- False Reject Rate (FRR)

- Depend on threshold: FAR down = FRR up

Abstracting from threshold
- FAR vs FRR plot (e.g. ROC curve)
- Equal Error Rate (EER), i.e. FAR = FRR
- FAR when FRR = 0 (no false accusations)
- FRR when FAR = 0 (no perp unpunished)

The Task: Test Corpus

Corpus:
- 8 students (Dutch)
- 9 texts from each student
  - fixed subjects
  - 3 argumentative, 3 descriptive, 3 fiction
  - about 1000 words per text
  - produced in controlled environment

Train: all texts with subject ≠ S
Test: all texts with subject S

Linguistic Profiling

General idea:

- make a profile (like a fingerprint) of the student's language use

- (check if it is distinguishing enough)

- measure any new text against the profile

Profiling with Lexical Features

```
Dit          #H#Dit         #H#Pron(aanw,neut)
is           #H#is          #H#V(ott,3,ev)-Misc(vreemd)
een          #H#een         #H#Art(onbep,zijdofonzijd,neut)-N
aspect       #H#aspect      #H#N(ev,neut)
van          #H#van         #H#Prep-N(ev,neut)-Adv(stell,onve
de           #H#de          #H#Art(bep,zijdofmv,neut)-N(ev,ne
Europese     #H#Europese    #H#Adj(stell,vervneut)-
    N(ev,neut)
eenwording #L#10+/L/ing #L#N(ev,neut)
.            #H#.           #H#Punc(punt)
```

Uni-, bi, trigrams of all combinations, e.g.

```
PCP=#H#is+#H#N(ev,neut)+#H#van
```

Sentence lengths, exact and grouped

```
LEN=9
LEN=1-10
```

Profiling with Lexical Features

Profile includes counts for:

- sentence lengths
- words / word patterns / word classes
- bi- and trigrams of above
- (single text occurrences filtered out)

Vector of about 100K counts

Counts are:

- normalized for text length
- expressed as relative under- or overuse

Profiling with Lexical Features

Author profile =
mean of the profiles for the known texts

Text verification score =
distance measure text profile to author profile

Profiling with Lexical Features

Distance measure:

$$\left(\sum | T_i - P_i |^{D} \; | T_i |^{S} \right)^{1/(D+S)} - \left(\sum | T_i |^{(D+S)} \right)^{1/(D+S)}$$

Orthogonalized:

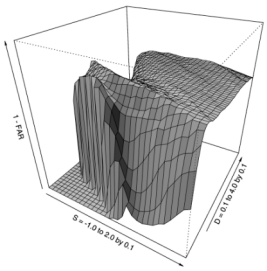
$$- \text{Mean}_{(\text{other author texts})}$$

$$/ \text{StdDev}_{(\text{other author texts})}$$

Results with Lexical Features

$FAR_{FRR=0}$
as function of D
and S

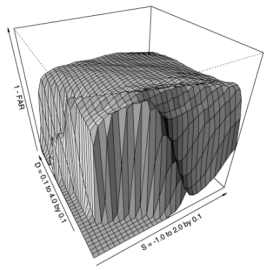
Best result 15%
(at D=0.60, S=0.15)



Results with Lexical Features

$FAR_{FRR=0}$
as function of D
and S

Best result 15%
(at D=0.60, S=0.15)



Profiling with Syntactic Features

Parse all texts (Amazon parser)
and extract all rewrites

Profile includes counts for:
- LHS label (constituent occurrence)
- LHS-RHS combos (dominance relations)
- LHS-RHS-RHS combos (linear precedence)

Vector of about 900K counts

Results with Syntactic Features

Parameter space not explored completely
(i.e. no nice picture)

Best result so far 25% (at D=1.3, S=1.4)

So is syntax useless?

Results with Syntactic Features

Parameter space not explored completely
(i.e. no nice picture)

Best result so far 25% (at D=1.3, S=1.4)

So is syntax useless?

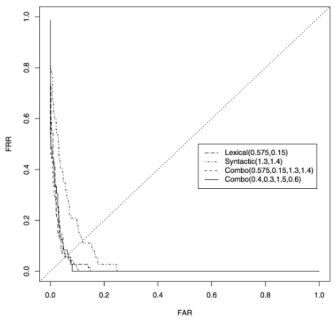
NO: combine lexical and syntactic

Results with Combination

Combination
= Addition

Combo best: 10%
Best combo: 8%

But see ROC and EER

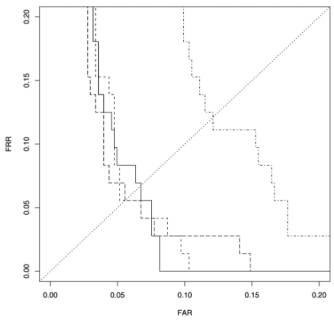


Results with Combination

Combination
= Addition

Combo best: 10%
Best combo: 8%

But see ROC and EER



Problem: Parameter Settings

So far, no automatic parameter selection!
(results above always best results)

Potential for improvement:

	LEX	SYN	COMB
Scenario above: Single threshold	14.9	24.8	8.1
Using fact that 7 vs 1: Renormalization	9.3	6.0	2.4
Optimal threshold: Oracle	0.8	1.6	0.2

Author Recognition and Sorting

	2-way errors/504	2-way % correct	8-way errors/72	8-way % correct
50 function w., PCA		c. 50%		
+ LDA		c. 60%		
+ entropy weighting		c. 80%		
All tokens, WPDV		97.8%		
LEX	6	98.8%	5	93%
SYN	14	98.2%	10	86%
COMB	3	99.4%	2	97%
LEX, renorm	1	99.8%	1	99%
SYN, renorm	4	99.2%	3	96%
COMB, renorm	0	100.0%	0	100%

Conclusion

- Upside
  - Linguistic Profiling viable
  - Improvements expected through
    - Better automatic parametrization
    - Larger amounts of text (of same type)
    - Further profiling features
- Downside
  - Needs substantial text base to start with
  - Need to find automatic parametrization

Final verdict:

YES, useful for this and other tasks

Part 3: Language Verification

If time left:

Back to: How good is this English?

Cf. COLING-2004 paper