Detection of Plagiarism and Text Reuse

Alberto Barrón-Cedeño and Paolo Rosso
http://www.dsic.upv.es/grupos/nle

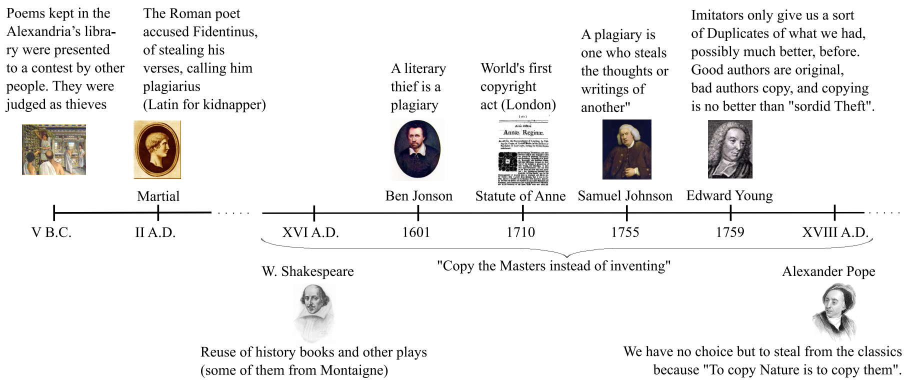
Natural Language Engineering Lab, ELiRF
Departmento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia, Spain

ICON 2010 Tutorial
Kharagpur, India
December 11th 2010

Outline

Introduction: Commercial Plagiarism Detection

Introduction: A "History" of Plagiarism



[Irribarne and Retondo, 1981, Lynch, 2006]

Introduction: In the news

JK Rowling sued for $\pounds 500m$ in plagiarism lawsuit by family of late Willy The Wizard author

16th June, 2009

George Harrison controversy vs The Chiffons for “My Sweet Lord”

1971

A Murcian professor is charged for plagiarising his student thesis

January 29th, 2009

The magistrate opens trial against Planeta for alleged plagiarism by Camilo José Cela

October 17th, 2010

Introduction: Plagiarism of Ideas

JK Rowling sued for $\pounds 500m$ in plagiarism lawsuit by family of late Willy The Wizard author

“Adrian Jacobs […] allegedly sent the manuscript to C. Little, the literary agent at Bloomsbury Publishing who went on to represent Miss Rowling, but it was rejected”

The magistrate opens trial against Planeta for alleged plagiarism by Camilo José Cela

…“given the coincidences in both books, La Cruz de San Andrés could be a partial plagiarism from ‘Carmen, Carmela, Carmiña’, written by María del Carmen Formoso Lapido, ”

- The narrative and events occurred in the books resemble each other. However, if plagiarism exists, it is of ideas (no words dependency)
- Plagiarism of ideas is nowadays (practically) impossible to be detected automatically

Introduction: Cut and Paste

A Murcian professor is charged for plagiarising his student’s thesis

A Valencian publisher edited the copied book

January 29th, 2009

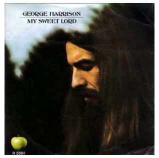
- It can be considered cut-and-paste plagiarism
- It is the easiest to detect

Introduction: Cryptomnesia

George Harrison vs The Chiffons

Music experts determined that “My Sweet Lord” was very similar to “He’s So Fine”, by Ronald Mack, played by The Chiffons (1962)

1971



- Plagiarism may occur in music, photography, painting and any other human made artifact (not only in text)

Cryptomnesia can give rise to unintended plagiarism, especially when logical memories are no longer recognised as memories, but are experienced as newly created ideas [Taylor, 1965]

Introduction: Plagiarism Definitions

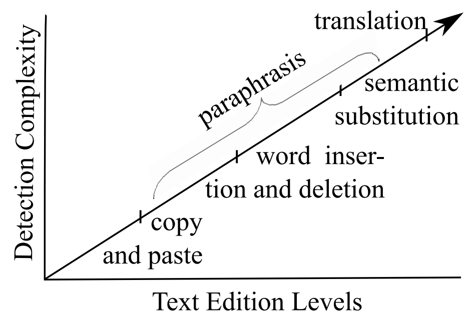
- to steal and pass off the ideas or words of another as one's own
- the reuse of someone else's prior ideas, processes, results, or words without explicitly acknowledging the original author and source
- giving incorrect information about the source of a quotation
- to take the thought or style of another writer whom one has never, never read

(from www.plagiarism.org, Merriam-Webster, IEEE and Devil's Dictionary)

Introduction: Plagiarism Commitment

- copy-paste
- paraphrasing
- idea plagiarism
- code plagiarism
- translated plagiarism

[Maurer et al., 2006]



Introduction: Is plagiarism?

- $\mathcal{A}$ Copying words or ideas from someone else without giving credit
- $\mathcal{A}'_1$ Copying the words and ideas from someone else's text without giving credit
- $\mathcal{A}'_2$ Changing words but copying the sentence structure of a source without giving credit
- $\mathcal{A}'_3$ Copiar las palabras o ideas de alguien más sin darle crédito

Introduction: is plagiarism?

- $\mathcal{A}$ Copying words or ideas from someone else without giving credit.
- $\mathcal{A}'_1$ Copying the words and ideas from someone else's text without giving credit.
- $\mathcal{A}'_2$ Changing words but copying the sentence structure of a source without giving credit.
- $\mathcal{A}'_3$ Copiar las palabras o ideas de alguien más sin darle crédito

$\mathcal{A}'_1$ is plagiarised. $\mathcal{A}'_2$ is not. $\mathcal{A}'_3$ is cross-language plagiarism

Introduction: Why is plag. detection interesting?

- Plagiarism is considered as one of the biggest problems in publishing, science, and education
- Text plagiarism is observed at an unprecedented scale with the advent of the World Wide Web (billions of texts, source codes, images, sounds, and videos easily accessible)
- The manual analysis of text with respect to plagiarism becomes infeasible on a large scale
- Plagiarism detection, the automatic identification of plagiarism and the retrieval of the original sources, is researched and developed as a possible countermeasure to plagiarism

Introduction: Copy-Paste Syndrome

- Today texts can be easily found, manipulated and combined
- The large amount of information resources, as digital libraries and the Web, have arisen new phenomena such as the so-called copy-paste syndrome
- Therefore, plagiarism has increased in recent years, which causes manual plagiarism detection infeasible

[Weber, 2007, Kulathuramaiyer and Maurer, 2007]

- New terms: cyberplagiarism [Comas and Sureda, 2008]

Introduction: Plagiarism and Text Reuse

Text reuse The activity whereby pre-existing written texts are used again to create a new text or version [Clough and Gaizauskas, 2009]

Plagiarism The reuse of someone else's prior ideas, processes, results, or words without explicitly acknowledging the original author and source [IEEE, 2008]

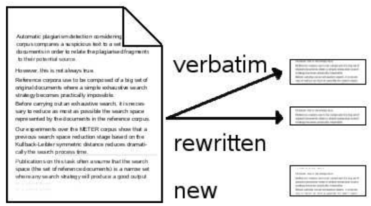
Text reuse	Plagiarism
newspapers	students reports
Wikis (Wikipedia)	web contents
collaborative authoring	scientific papers

Automatic plagiarism detection assists the human.

Introduction: METER project

- Compiled with journalists
- News provided by the Press Association
- Versions of the same news published by 9 newspapers

[Clough et al., 2002]



Introduction: The METER Project

PA version	The Telegraph version
Celebrity chef Marco Pierre White today won the battle of the Titanic and Atlantic restaurants. Oliver Peyton, owner of the Atlantic Bar and Grill, had tried to sink Marco's new Titanic restaurant housed in the same West End hotel in London by seeking damages against landlords Forte Hotels and an injunction in the High Court. But today the Atlantic announced in court it had reached a confidential agreement with the landlords and was discontinuing the whole action.	THE chef Marco Pierre White yesterday won a dispute over the Titanic and Atlantic restaurants. Oliver Peyton, owner of the Atlantic, had tried to close White's new Titanic restaurant, housed in the same West End hotel in London, by seeking damages against the landlords, Forte Hotels, and a High Court injunction. He claimed that the Titanic was a replica of the Atlantic and should not be allowed to trade in competition at the Regent Palace Hotel.

Introduction: The METER Project

PA version	The Telegraph version
Celebrity chef Marco Pierre White today won the battle of the Titanic and Atlantic restaurants. Oliver Peyton, owner of the Atlantic Bar and Grill, had tried to sink Marco's new Titanic restaurant housed in the same West End hotel in London by seeking damages against landlords Forte Hotels and an injunction in the High Court. But today the Atlantic announced in court it had reached a confidential agreement with the landlords and was discontinuing the whole action.	THE chef Marco Pierre White yesterday won a dispute over the Titanic and Atlantic restaurants. Oliver Peyton, owner of the Atlantic, had tried to close White's new Titanic restaurant, housed in the same West End hotel in London, by seeking damages against the landlords, Forte Hotels, and a High Court injunction. He claimed that the Titanic was a replica of the Atlantic and should not be allowed to trade in competition at the Regent Palace Hotel.

Activity 1: Detecting text reuse over the METER corpus

Introduction: The METER Project

Feature	Value
Reference corpus size (kb)	1,311
Number of PA notes	771
Tokens / Types	226k / 25k
Suspicious corpus size (kb)	828
Number of newspapers notes	444
Tokens / Types	139k / 19k
Entire corpus tokens	366k
Entire corpus types	33k

Introduction: Plagiarism Detection Task

Given a (set of) suspicious document(s) and a set of source documents, find all plagiarised sections in the suspicious document(s) and, if available, the corresponding source sections.

Afterwards, a person can take the final decision: whether a text has been reused or not and if it is plagiarised.

Introduction: Plagiarism Analysis Taxonomy

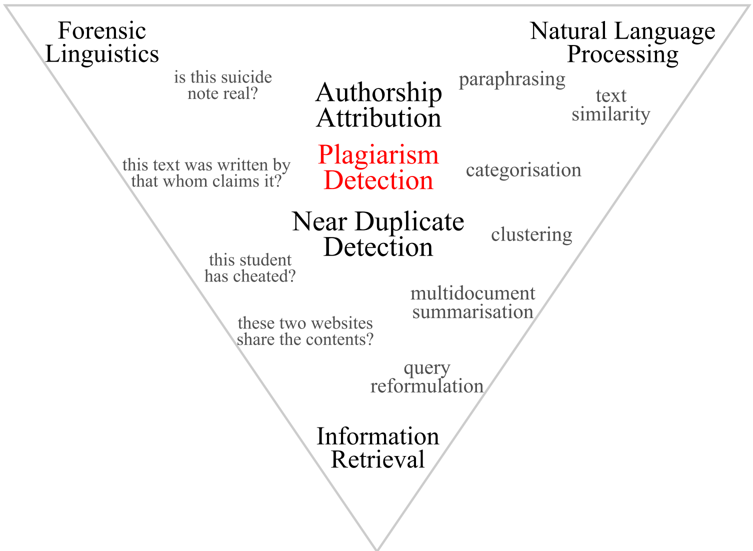
- Plagiarism type / Detection principle
  - Exact
    - Large part of document / Document model comparison (suffix-tree)
    - Small part of document / Local identity analysis
      - Chunk identity analysis (with reference)
      - Style analysis (without reference)
  - Modified
    - Language translation / Cross-language similarity analysis
    - Reformulation / Similarity analysis
      - Large part of document / Document model comparison (VSM)
      - Small part of document / Local similarity analysis
        - Fingerprinting (with reference)
        - Style analysis (without reference)

[Meyer zu Eißen and Stein, 2006]

Introduction: Drawbacks

1. plagiarism implies an infringement and, due to ethical aspects, no standard collection of real plagiarism cases is available;
2. the source of a plagiarism may be hosted on large collections of documents (sometimes forgotten by researchers);
3. plagiarism often implies modifications such as words substitution, paraphrasing, and even translation.

Introduction: Location of the Problem



Outline

Basics: n-grams

An n-gram is a sequence of overlapping units of length n over a given sample (characters, words, sounds, etc).

- character 3-grams

example

exa xam
amp mpl
ple

- word 2-grams

this is just an example

this is
is just
just an
an example

Basics: Hash Function

"any well-defined procedure or mathematical function that converts a large, possibly variable-sized amount of data into a small datum, [...] that may serve as an index to an array. The values returned by a hash function are called hash values, hash codes, hash sums, checksums or simply hashes." [Wikipedia, 2010a]

For instance:

- $md5sum$(this is a test) $= e19c1283c925b3206685ff522acfe3e6$
- $RabinKarp$(starwarsisanepicspaceoperafranchiseinitiallyconcei) $=$ 4742204955

The probability of collision is extremely low.

Basics: Text complexity

Gunning fog index

$$I_G = 0.4\left(\frac{|words|}{|sentences|} + 100 * \frac{|complex\ words|}{|words|}\right)$$

(complex words are those with three or more syllables)

$$\begin{aligned} I_G(comic) &= 6 \\ I_G(Newsweek) &= 10 \\ I_G(T_1) &= 15.2 \\ I_G(T_2) &= 14.1 \end{aligned}$$

(also Flesch–Kincaid readability test, among others)

Basics: Word Frequency Class

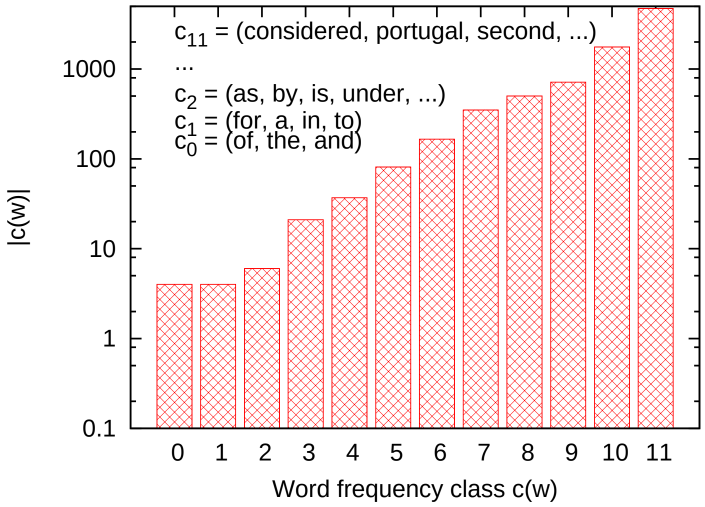
Given the corpus $\mathcal{D}$, the word frequency class is defined as:

$$c(w) = \lfloor log_2(f(w^*)/f(w)) \rfloor$$

where w^* is the most frequently used word in $\mathcal{D}$

	w	$f(w)$	$c(w)$
w^*	the	6,047,424	0
	of	2,887,888	1
	and	2,615,135	1
	house	49,295	6
	undertaken	2,699	11
	corpus	723	13

Basics: Word Frequency Class



Basics: Text similarity

Relevance of Text Similarity Estimation

- Information flow tracking [Metzler et al., 2005]
- Clustering and categorisation [Bigi, 2003]
- Multi-document summarisation [Goldstein et al., 2000]
- Version control [Hoad and Zobel, 2003]
- Text re-use analysis [Clough et al., 2002]
- Plagiarism detection [Maurer et al., 2006]

Basics: Similarity Measures

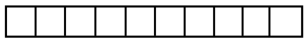
$$sim(d, d_q) \in [0, 1]$$

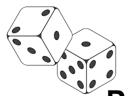
- $sim(d, d_q) = 0 \rightarrow d$ and d_q are not similar at all
- $sim(d, d_q) = 1 \rightarrow d$ and d_q are highly similar

However, note that such optimal measures are not always at hand.

Basics: Similarity Measures

Jaccard coefficient
Cosine similarity
Word chunking overlap
Vector Space





Fingerprinting
Winnowing
SPEX

Probabilistic
Machine Translation
Kullback-Leibler
Okapi BM25

Basics: Similarity Measures Illustration

Wikipedia article “Star Wars”

d	star wars is an epic space opera franchise initially conceived by george lucas during the 1970s and significantly expanded since that time . the first film in the franchise was simply titled star wars , but later had the subtitle a new hope added to distinguish it from its sequels and prequels .
d'	star wars is an epic space opera franchise initially conceived by george lucas . the first film in the franchise was simply titled star wars , but later had the subtitle episodeiv : a new hope added to distinguish it from its sequels and prequels .

Basics: Similarity Measures - VSM

Jaccard Coefficient

$$\omega_t = \{0, 1\}$$

$$sim(d, d_q) = J(d, d_q) = \frac{|v_d \cap v_{d_q}|}{|v_d \cup v_{d_q}|}$$

[Jaccard, 1901]

Basics: Similarity Measures - VSM

Jaccard Coefficient

d

,	epic	it	star
.	expanded	its	subtitle
1970s	film	later	that
a	first	lucas	the
added	franchise	new	time
an	from	opera	titled
and	george	prequels	to
but	had	sequels	wars
by	hope	significantly	was
conceived	in	simply	
distinguish	initially	since	
during	is	space	

d'

,	distinguish	in	sequels
:	epic	initially	simply
.	episodeiv	is	space
a	film	it	star
added	first	its	subtitle
an	franchise	later	the
and	from	lucas	titled
but	george	new	to
by	had	opera	wars
conceived	hope	prequels	was

$$sim(d, d_q) = J(d, d_q) = \frac{|v_d \cap v_{d_q}|}{|v_d \cup v_{d_q}|} = 0.7916$$

Basics: Similarity Measures - VSM

Cosine Similarity

$$\omega_t \in [0, 1]$$

ω_t is estimated by the well known tf

$$sim(d, d_q) = \frac{\sum_{t \in d \cap d_q} \left(\omega_{t,d} \cdot \omega_{t,d_q}\right)}{\sqrt{\sum_{t \in d} \left(\omega_{t,d}\right)^2 \cdot \sum_{t_q \in d_q} \left(\omega_{t,d_q}\right)^2}}$$

Basics: Similarity Measures - VSM

Cosine Similarity

d

,	1	first	1	prequels	1
.	3	franchise	2	sequels	1
1970s	1	from	1	significantly	1
a	1	george	1	simply	1
added	1	had	1	since	1
an	1	hope	1	space	1
and	2	in	1	star	1
but	1	initially	1	subtitle	1
by	1	is	1	that	1
conceived	1	it	1	the	4
distinguish	1	its	1	time	1
during	1	later	1	titled	1
epic	1	lucas	1	to	1
expanded	1	new	1	wars	2
film	1	opera	1	was	1

d'

,	1	film	1	lucas	1
:	1	first	1	new	1
.	2	franchise	2	opera	1
a	1	from	1	prequels	1
added	1	george	1	sequels	1
an	1	had	1	simply	1
and	1	hope	1	space	1
but	1	in	1	star	2
by	1	initially	1	subtitle	1
conceived	1	is	1	the	3
distinguish	1	it	1	titled	1
epic	1	its	1	to	1
episodeiv	1	later	1	wars	2
				was	1

$$sim(d, d_q) = \frac{\sum_{t \in d \cap d_q} \left(\omega_{t,d} \cdot \omega_{t,d_q}\right)}{\sqrt{\sum_{t \in d} \left(\omega_{t,d}\right)^2 \cdot \sum_{t_q \in d_q} \left(\omega_{t,d_q}\right)^2}} = 0.9242$$

Basics: Similarity Measures - VSM

Word Chunking Overlap

$$\omega_t \in [0, 1]$$

- Based on the so called asymmetric subset measure:

$$subset(d, d') = \frac{\sum_{t_i \in c(d,d')} tf_{t,d} \cdot tf_{t,d'}}{\sum_{t_i \in d} tf_{t_i,d}^2}$$

  - $c(d, d_q)$ is a closeness set containing those terms $t \in d \cap d_q$ matching the condition $tf_{t,d} \sim tf_{t,d_q}$. t belongs to $c(d, d_q)$ if:

$$\varepsilon - \left(\frac{tf_{t,d}}{tf_{t,d'}} + \frac{tf_{t,d'}}{tf_{t,d}}\right) > 0$$

[Shivakumar and García-Molina, 1995]

Basics: Similarity Measures - VSM

Word Chunking Overlap

- ε defines how close the frequency of t in both documents must be in order to be included in the closeness set (for instance, $\varepsilon = 2.5$)

$$sim'(d, d_q) = max\left\{subset(d, d_q), subset(d_q, d)\right\}$$

As $sim'(d, d_q)$ may be higher than 1, it can be normalised to fit the range $[0, 1]$:

$$sim(d, d_q) = \frac{sim'(d, d_q)}{\max_{d' \in D} sim'(d', d_q)}$$

[Shivakumar and García-Molina, 1995]

Basics: Similarity Measures - VSM

Word Chunking Overlap

- By considering $\varepsilon = 2.5$

d

,	1	franchise	2	opera	1
.	3	from	1	prequels	1
a	1	george	1	sequels	1
added	1	had	1	simply	1
an	1	hope	1	space	1
and	2	in	1	star	1
but	1	initially	1	subtitle	1
by	1	is	1	the	4
conceived	1	it	1	titled	1
distinguish	1	its	1	to	1
epic	1	later	1	wars	2
film	1	lucas	1	was	1
first	1	new	1		

d'

,	1	franchise	2	opera	1
.	2	from	1	prequels	1
a	1	george	1	sequels	1
added	1	had	1	simply	1
an	1	hope	1	space	1
and	1	in	1	star	1
but	1	initially	1	subtitle	1
by	1	is	1	the	3
conceived	1	it	1	titled	1
distinguish	1	its	1	to	1
epic	1	later	1	wars	2
film	1	lucas	1	was	1
first	1	new	1		

$$sim'(d, d_q) = max\left\{0.8857, 1.0689\right\}$$

$$sim(d, d_q) = \frac{1.0689}{max_{d' \in D} sim'(d', d_q)}$$

Basics: Similarity Measures - Fingerprinting

- A family of models designed to efficiently compare texts
- Documents are sub-sampled
- Samples are codified as hashes: $d \rightarrow H^*_d$
- The hashes compose the fingerprint

Basics: Similarity Measures - Fingerprinting

Winnowing

- It considers character-level q-grams
- Based on the selection of chunks obtained by a sliding window passing over the text
- Parameters:
  1. $q = 50$ (noise threshold). It defines the level of the q-grams
  2. $t = 100$ (guarantee threshold). It defines the length of the sliding window.
- The lowest hash values of each window compose the fingerprint

[Schleimer et al., 2003]

Basics: Similarity Measures - Fingerprinting

Winnowing

d	d'
starwarsisanepicspaceoperafra nchiseinitiallyconceivedbygeorg elucasduringthe1970	starwarsisanepicspaceope rafranchiseinitiallyconceive dbygeorgelucas
[4742204955 4690954177 51549901 624610790 -2470793273 [-1315199375 3953400264 -78415511 [664863318 3374288481] 4230663014 -3213422081 -2056259009 7513105677 -6553730326] 5257922027 4828416784 -8476824670] 9011767372 1240867252]	[4742204955 4690954177 51549901 624610790 -2470793273 -1315199375 3953400264]

By considering $t = 20$ $$sim(d, d_q) = \frac{\emptyset}{2} = 0$$

By considering $t = 10$ $$sim(d, d_q) = \frac{1}{3}$$

Basics: Similarity Measures - Fingerprinting

SPEX

- word-level chunks
- "if any sub-chunk of any chunk can be shown to be unique, then the chunk in its entirety must be unique"
- Hashes occurring in only one document are not relevant.
- Given D, the task is to identify those chunks appearing in more than one document $d \in D$. The main steps are:
  1. To generate a list h_1 of 1-grams over D and to count in how many documents each of them occur.
  2. In the next steps h_n is built by selecting only those n-grams g fulfilling the condition that h_{n-1} contains $g_{[0,n-1]}$ and $g_{[1,n]}$ and both are counted two times ($\text{máx}(n) = 8$).

[Bernstein and Zobel, 2004]

Basics: Similarity Measures - Fingerprinting

SPEX

$$sim(d, d_q) = \frac{1}{mean(|d|, |d_q|)} \sum_{c \in d \wedge c \in d_q} 1$$

where $mean(|d|, |d_q|)$ is the mean length of the documents d and d_q.

[Bernstein and Zobel, 2004]

Basics: Similarity Measures - Fingerprinting

SPEX

$n=1$, d:

star	significantly	later
wars	expanded	had
is	since	the
an	that	subtitle
epic	time	a
space	the	new
opera	first	hope
franchise	film	
initially	in	
conceived	the	
by	franchise	
george	was	
lucas	simply	
during	titled	
the	star	
1970s	wars	
and	but	

d':

star	first	subtitle
wars	film	episodeiv
is	in	a
an	the	new
epic	franchise	hope
space	was	
opera	simply	
franchise	titled	
initially	star	
conceived	wars	
by	but	
george	later	
lucas	had	
the	the	

Basics: Similarity Measures - Fingerprinting

SPEX

$n=2$, d:

star wars	in the
wars is	the franchise
is an	franchise was
an epic	was simply
epic space	simply titled
space opera	titled star
opera franchise	star wars
franchise initially	wars but
initially conceived	but later
conceived by	later had
by george	had the
george lucas	the subtitle
the first	subtitle a
first film	a new
film in	new hope

d':

star wars	in the
wars is	the franchise
is an	franchise was
an epic	was simply
epic space	simply titled
space opera	titled star
opera franchise	star wars
franchise initially	wars but
initially conceived	but later
conceived by	later had
by george	had the
george lucas	the subtitle
lucas the	subtitle a
the first	a new
first film	new hope
film in	

Basics: Similarity Measures - Fingerprinting

SPEX

$n=3$, d:

star wars is	in the franchise
wars is an	the franchise was
is an epic	franchise was s…
an epic space	was simply titled
epic space opera	simply titled star
space opera fran…	titled star wars
opera franchise in…	star wars but
franchise initially …	wars but later
initially conceived …	but later had
conceived by g…	later had the
by george lucas	had the subtitle
the first film	the subtitle a
first film in	subtitle a new
film in the	a new hope

d':

star wars is	in the franchise
wars is an	the franchise was
is an epic	franchise was s…
an epic space	was simply titled
epic space opera	simply titled star
space opera fra…	titled star wars
opera franchise in…	star wars but
franchise initially …	wars but later
initially conceived …	but later had
conceived by g…	later had the
by george lucas	had the subtitle
the first film	the subtitle a
first film in	subtitle a new
film in the	a new hope

Basics: Similarity Measures - Fingerprinting

SPEX

- By considering $l = 3$ (higher could be better)

$$sim(d, d_q) = \frac{1}{mean(|d|, |d_q|)} \sum_{c \in d \wedge c \in d_q} 1$$

$$sim(d, d_q) = \frac{1}{49.5} \cdot 28 = 0.56$$

Basics: Similarity Measures - Probabilistic

- d is characterised by the probability associated to its tokens
- $sim(d, d_q)$ can be approached by calculating the probability of their relation.
- The output of these models is not ranged in $[0, 1]$)

Basics: Similarity Measures - Probabilistic

Machine Translation

- Given a text e written in a language L, to find the most likely translation f, in a language L'
- Adaptation of the IBM Model 1 [Brown et al., 1993]. by considering $L = L'$ [Berger and Lafferty, 1999, Metzler et al., 2005]

Basics: Similarity Measures - Probabilistic

Machine Translation. IBM Model Adaptation

$$sim(d, d_q) = \varrho(d)\, w(d_q \mid d)$$

- $\varrho(d)$ is a length model probability (as $L = L'$, $\varrho(d) = 1$)
- $w(d_q \mid d)$ is a tailored version of the translation model probability:

$$w(d_q \mid d) = \prod_{x \in d_q} \sum_{y \in d} p(x, y)$$

  - $p(x, y)$ is a dictionary containing the probability that word x is a translation of word y: $p(x, y) = 1$ if $x = y$ and 0 otherwise.

Basics: Similarity Measures - Probabilistic

Machine Translation. IBM Model Adaptation

- In order to handle entire documents.

$$w(d_q \mid d) = \sum_{x \in d_q} \sum_{y \in d} p(x, y)$$

For each word $x \in d_q \setminus d$, a penalisation $\varepsilon = -0.1$ may be applied

$$sim(d, d_q) = \frac{sim'(d, d_q)}{\text{máx}_{d' \in D}\, sim'(d', d_q)}$$

Basics: Similarity Measures - Probabilistic

Machine Translation

$n = 1$	d	
star	during	franchise
wars	the	was
is	1970s	simply
an	and	titled
epic	significantly	star
space	expanded	wars
opera	since	but
franchise	that	later
initially	time	had
conceived	the	the
by	first	subtitle
george	film	a
lucas	in	new
	the	hope

$$w(d_q \mid d) = 33 - 0.8 = 32.2$$

	d'	
star	george	star
wars	lucas	wars
is	the	but
an	first	later
epic	film	had
space	in	the
opera	the	subtitle
franchise	franchise	episodeiv
initially	was	a
conceived	simply	new
by	titled	hope

$$sim(d, d_q) = \frac{32.2}{max_{d' \in D} sim'(d', d_q)}$$

Basics: Similarity Measures - Probabilistic

Kullback-Leibler distance

- KL_δ is a symmetric version of the Kullback-Leibler Divergence [Kullback and Leibler, 1951].
- It measures how close two probability distributions P and Q are

$$KL_\delta(P_{d_q} \mid\mid Q_d) = \sum_{x \in \mathcal{X}} (P(x) - Q(x)) log \frac{P(x)}{Q(x)}$$

- P_{d_q} and Q_d are distributions of tokens
- P_{d_q} is composed of the top 20 % of the terms in d_q ranked by tf-idf
- Q_d is composed of the same terms of P_{d_q} after a smoothing process

Basics: Similarity Measures - Probabilistic

Kullback-Leibler distance

- KL measures the distance instead of the similarity
- $KL_\delta(P_{d_q} \mid\mid Q_d) = 0 \rightarrow P_{d_q} = Q_d$ and the documents are quite similar.

$$sim(d, d_q) = -\left(\frac{KL_\delta(P_{d_q} \mid\mid Q_d)}{max_{d'} KL(P_{d_q} \mid\mid Q_d)} - 1 \right)$$

Basics: Similarity Measures - Probabilistic

Kullback-Leibler

keywords in d ranked by tf-idf		
1970s	expanded	new
lucas	titled	but
george	conceived	was
star	initially	had
wars	film	is
epic	simply	that
franchise	during	an
subtitle	since	and
hope	later	by
opera	time	a
subtitle	space	in
hope	first	the

P_{d_q}	
1970s	0.01886
lucas	0.01886
george	0.01886
star	0.03773
wars	0.03773
epic	0.01886
franchise	0.03773

Q_d	
1970s	0.0002
lucas	0.0216
george	0.0216
star	0.0433
wars	0.0433
epic	0.0213
franchise	0.0433

Basics: Similarity Measures - Probabilistic

Kullback-Leibler distance

$$KL_\delta(P_{d_q} \mid\mid Q_d) = \sum_{x \in \mathcal{X}} (P(x) - Q(x)) log \frac{P(x)}{Q(x)} = 0.08817$$

$$sim(d, d_q) = -\left(\frac{0.08817}{max_{d'} KL(P_{d_q} \mid\mid Q_d)} - 1 \right)$$

Basics: Similarity Measures - Probabilistic

Okapi BM25

- It extends the approach of idf by additionally considering tf and document length [Spärck Jones et al., 2000]

$$BM25(d, d_q) = \sum_{t \in d_q} idf_t \cdot \alpha_{t,d} \cdot \beta_{t,d_q}$$

where

$$\alpha_{t,d} = \frac{(k_1 + 1)\, tf_{t,d}}{k_1 \left((1 - b) + b \cdot \frac{|d|}{L_{avg}} \right) + tf_{t,d}}$$

- $k_1 = 0$ corresponds to a binary model (not considering tf)
- $b = 0$ corresponds to no length normalisation; $b = 1$ corresponds to a full scaling of the term weight to the document length.
- For instance, $k_1 = 1.2$ and $b = 0.75$
- L_{avg} is the average document length in the collection

Basics: Similarity Measures - Probabilistic

Okapi BM25

- β_{t,d_q} normalises the tf of the terms in d_q:

$$\beta_{t,d_q} = \frac{(k_3 + 1)\, tf_{t,d_q}}{k_3 + tf_{t,d_q}}$$

- $k_3 = 2$. k_1 of α and k_3 of β are calibrators of the tf.

$$sim(d, d_q) = \frac{sim'(d, d_q)}{\text{máx}_{d' \in D}\, sim'(d', d_q)}$$

Outline

Intrinsic Plagiarism Detection

An expert is often able to detect plagiarism by reading a document

Insertion of text from a different author into d_q causes style and complexity irregularities

Quantification can be made by measuring…

Text readability	Gunning Fog, Flesch–Kincaid
Vocabulary richness	types/tokens ratio
Basic statistics	avg. sentence length, avg. word length
n-grams profiles	character level statistics

[Meyer zu Eißen and Stein, 2006, Stamatatos, 2009]

Intrinsic Plagiarism Detection

En este trabajo, hemos hecho una investigación acerca de la influencia que tiene la cantidad de sales minerales en el humor de las personas. Para la investigación he trabajado con 5 personas que han tomado agua con distinta cantidad de sales minerales. Nuestra teoría es que entre más sales minerales haya en el agua, las personas son más volubles. […]
Las sales minerales son moléculas inorgánicas de fácil ionización en presencia de agua y que en los seres vivos aparecen tanto precipitadas como disueltas. Las sales minerales disueltas en agua siempre están ionizadas. Estas sales tienen función estructural y funciones de regulación del pH, de la presión osmótica y de reacciones bioquímicas, en las que intervienen iones específicos. […]
Me parece que los resultados son buenos. […]

Intrinsic Plagiarism Detection

En este trabajo, hemos hecho una investigación acerca de la influencia que tiene la cantidad de sales minerales en el humor de las personas. Para la investigación he trabajado con 5 personas que han tomado agua con distinta cantidad de sales minerales. Nuestra teoría es que entre más sales minerales haya en el agua, las personas son más volubles. […]
Las sales minerales son moléculas inorgánicas de fácil ionización en presencia de agua y que en los seres vivos aparecen tanto precipitadas como disueltas. Las sales minerales disueltas en agua siempre están ionizadas. Estas sales tienen función estructural y funciones de regulación del pH, de la presión osmótica y de reacciones bioquímicas, en las que intervienen iones específicos. […]
Me parece que los resultados son buenos. […]

Intrinsic Plagiarism Detection

En este trabajo, hemos hecho una investigación acerca de la influencia que tiene la cantidad de sales minerales en el humor de las personas. Para la investigación he trabajado con 5 personas que han tomado agua con distinta cantidad de sales minerales. Nuestra teoría es que entre más sales minerales haya en el agua, las personas son más volubles. [...]
Las sales minerales son moléculas inorgánicas de fácil ionización en presencia de agua y que en los seres vivos aparecen tanto precipitadas como disueltas. Las sales minerales disueltas en agua siempre están ionizadas. Estas sales tienen función estructural y funciones de regulación del pH, de la presión osmótica y de reacciones bioquímicas, en las que intervienen iones específicos. [...]
Me parece que los resultados son buenos. [...]

Intrinsic Plagiarism Detection

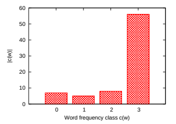
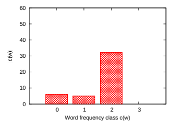
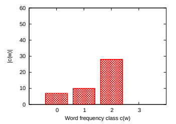
- Word average frequency class
- Average sentence length
- Average word length
- Stop-words average
- Complexity measures

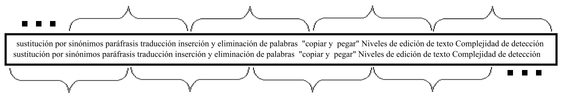
[Meyer zu Eißen and Stein, 2006]

Intrinsic Plagiarism Detection

En este trabajo, hemos hecho una investigación acerca de la influencia que tiene la cantidad de sales minerales en el humor de las personas. Para la investigación he trabajado con 5 personas que han tomado agua con distinta cantidad de sales minerales. Nuestra teoría es que entre más sales minerales haya en el agua, las personas son más volubls. [...]
Las sales minerales son moléculas inorgánicas de fácil ionización en presencia de agua y que en los seres vivos aparecen tanto precipitadas como disueltas. Las sales minerales disueltas en agua siempre están ionizadas. Estas sales tienen función estructural y funciones de regulación del pH, de la presión osmótica y de reacciones bioquímicas, en las que intervienen iones específicos. [...]
Me parece que los resultados son buenos. [...]

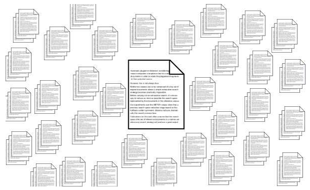
Intrinsic Plagiarism Detection

Measure	Global	■ (red)	■ (black)
tokens	135	63	72
types	78	44	46
W. avg. freq. class			
avg. sentence length	19.28	21.00	18.00
avg. word length	4.93	5.38	4.54
Complex. measures	16.72	17.07	13.82



Outline

External Plagiarism Detection



- Better evidence than style and complexity irregularities is if the source of plagiarism case can be provided
- It is closer to Information Retrieval

d_q and a collection of potential source documents D are given. The task is to identify the plagiarised sections in d_q (if there are any), and their respective source sections in D

[Potthast et al., 2009]

External Plagiarism Detection

Issues that render this task difficult

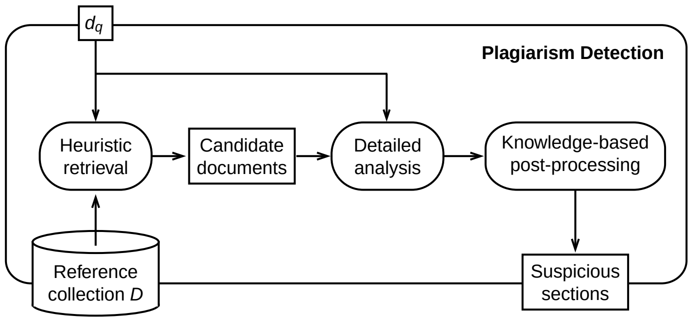
- Number of potential source documents, $|D|$;
- Plagiarising a text often includes paraphrasing, summarising, and even translation.

Models

Vector Space Models	[Broder, 1997], [Maurer et al., 2006]	
Fingerprinting techniques	SPEX	[Bernstein and Zobel, 2004]
	Winnowing	[Schleimer et al., 2003]

[Potthast et al., 2009]

External: Prototypical Process

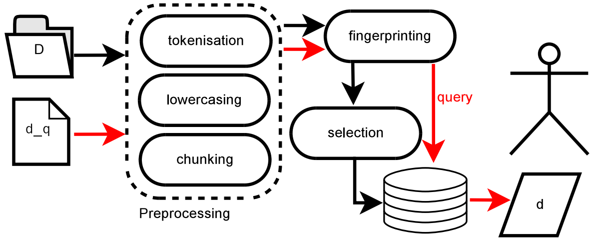


Adapted from [Stein et al., 2007]

External: Countermeasures

source Copying words or ideas from someone else without giving credit.

cut-and-paste Copying words or ideas from someone else without giving credit.



[Brin et al., 1995, Schleimer et al., 2003]

External: Fingerprinting (+ Winnowing)

COPS: COpy Protection System

- $\mathcal{A}$ creates a new work d and she registers it to a server
- d is broken into small units; sentences
- each sentence is hashed and a pointer to it is stored in a large hash table

[Brin et al., 1995]

External: Fingerprinting

COPS: COpy Protection System

given d':
break d' into chunks
for each chunk d'_i in d':
Calculate $\mathcal{H}(d'_i)$
Search for $\mathcal{H}(d'_i)$ into the data base

The amount of common words/sentences between d and d' is considered in order to decide whether they are related.

External: Fingerprinting

COPS: COpy Protection System

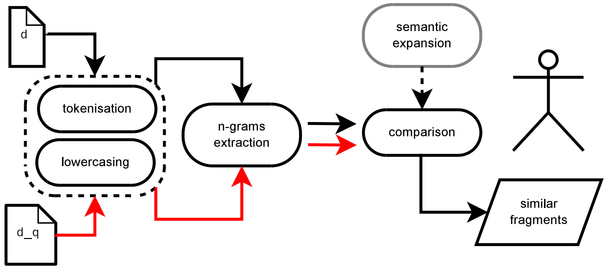
- "The electronic medium makes it much easier to illegally copy and distribute information"
- "one would like to have an infrastructure that gives users access to a wide variety of [...] information sources, but that at the same time gives information providers good economic incentives for offering their information"
- "users can be allowed to browse through low-resolution copies of documents, or through documents that have key components missing"

1995: a "classic model"

External: Countermeasures

source Copying words or ideas from someone else without giving credit.

modified copy Copying the words and ideas from someone else's text without giving credit.



[Broder, 1997, Kang et al., 2006]

External: n-grams

n-gram Based Detection

- $N(d)$ is the set of n-grams in $d \in D$
- $s \in S$ is split into sentences $s_{\{1 \ldots i \ldots I\}}$
- $N(s_i)$ is the set of n-grams in s_i
- The containment measure (cosine or Jaccard coefficient) can be calculated [Broder, 1997]

$$C(s_i \mid d) = \frac{|N(s_i) \cap N(d)|}{|N(s_i)|}$$

[Barrón-Cedeño and Rosso, 2009]

External: n-grams

Why n-grams work?

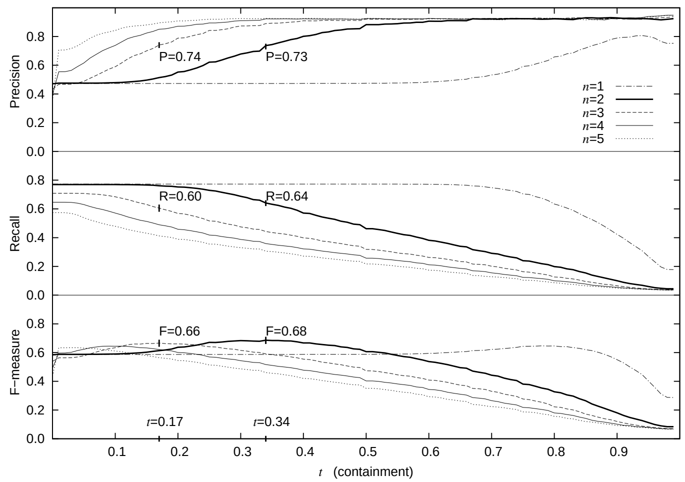
- 4 documents (3,728 words in average)
- One author $\mathcal{A}$
- One topic

Documents	1-grams	2-grams	3-grams	4-grams
2	0.1692	0.1125	0.0574	0.0312
3	0.0720	0.0302	0.0093	0.0027
4	0.0739	0.0166	0.0031	0.0004

Activity 2: Increase n until getting a hapax legomena on the Web

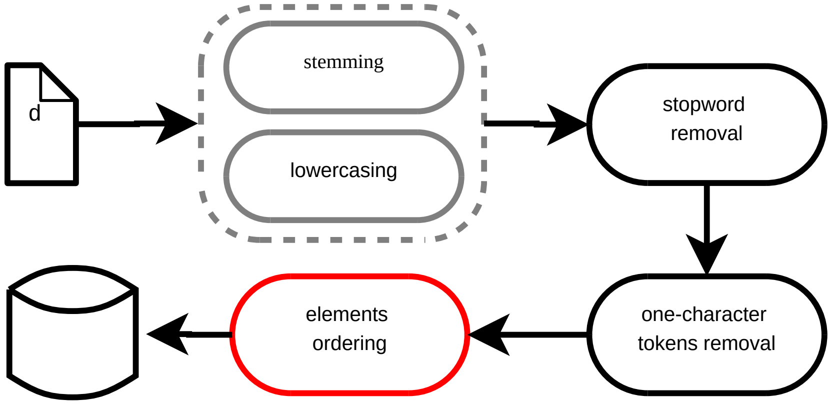
[Barrón Cedeño, 2008]

External: Definition of n (METER Corpus)



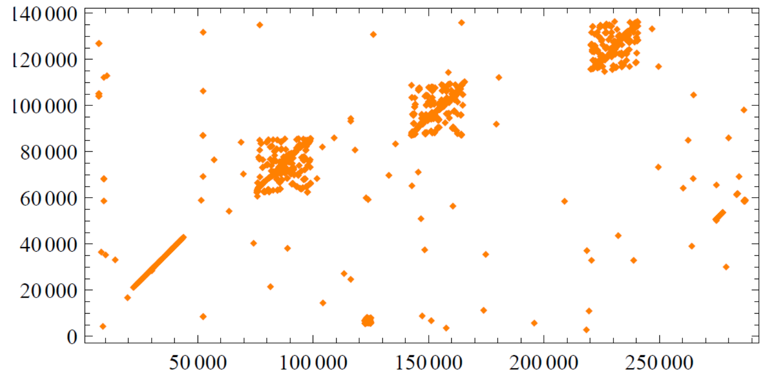
[Barrón-Cedeño and Rosso, 2009]

External: Contextual n-grams



[Rodríguez Torrejon and Martín Ramos, 2010a, Rodríguez Torrejon and Martín Ramos, 2010b]

External: Dotplot techniques



[Basile et al., 2009, Grozea et al., 2009]

External: Vocabulary Expansion

- Based on word comparison at sentence level
- Vocabulary expansion with Wordnet (Wikipedia is useful as well)

(Mark Haddon, 2003)

The curious incident of the dog in the night time

The peculiar incident of the cat in the late day time

[Kang et al., 2006]

External: Vocabulary Expansion

- Based on word comparison at sentence level
- Vocabulary expansion with Wordnet (Wikipedia is useful as well)

(Mark Haddon, 2003)

The curious incident of the dog in the night time

synonym antonym ~hypernym

The peculiar incident of the cat in the day time

[Kang et al., 2006]

External: Fuzzy Fingerprinting

- Fingerprint as an indicator for a high similarity between the fingerprinted objects
- The similarity between d_1 and d_2 is measured by a function $\varphi(\mathbf{d}_1, \mathbf{d}_2)$
- $\varphi(\mathbf{d}_1, \mathbf{d}_2)$ maps onto $[0, 1]$ (no and maximum similarity)

[Stein, 2005]

External: Fuzzy Fingerprinting

- The fuzzy hash function to compute the fingerprint $h_\varphi(d)$ is based on prefix frequency classes: c_a, c_b, c_c, …, c_z
- A standard distribution of index term frequencies can be stated (BNC)
- From a pre-defined set of prefixes, the a priori probability of a term being member in a prefix class can be stated
- The deviation of a document's term distribution from the a priori probabilities forms its fingerprint

External: Fuzzy Fingerprinting

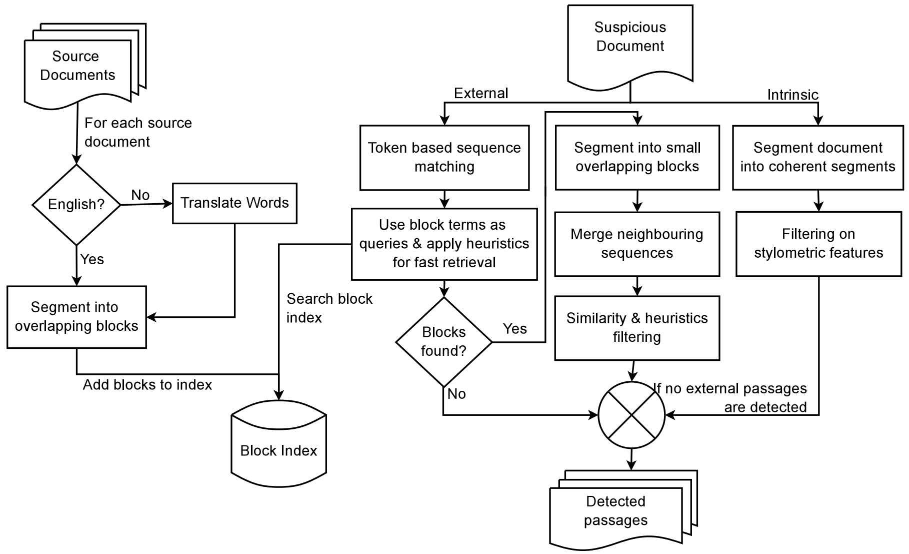
The fuzzy fingerprint $h_\varphi(d)$ is constructed within the following steps:

1. Extraction of the set of index terms from d
2. Computation of pf, the vector of relative frequencies of the prefix classes in d
3. Computation of Δ_{pf} (vector of deviations to the expected distribution)
4. Fuzzyfication of Δ_{pf}

Hash collision

$$h_\varphi(d) \cap h_\varphi(d') \neq \emptyset \Rightarrow \varphi\left(\mathbf{d}, \mathbf{d}'\right) \geq 1 - \varepsilon$$

External: IR Approach



[Muhr et al., 2010]

Outline

Introduction

Basic Concepts

Intrinsic Plagiarism Detection

External Plagiarism Detection

Cross-Language Plagiarism Detection

Plagiarism Detection Competition

Not Only Plain Text, Not only Plagiarism

Start Point

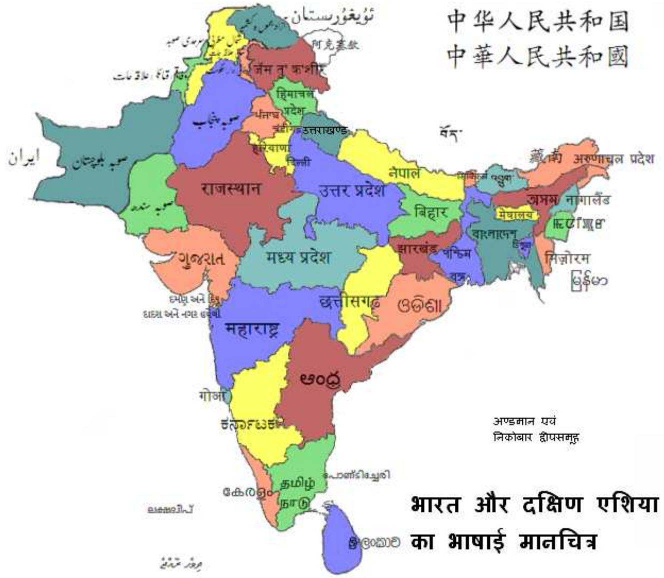
Cutting the Edge

CL Plagiarism Detection

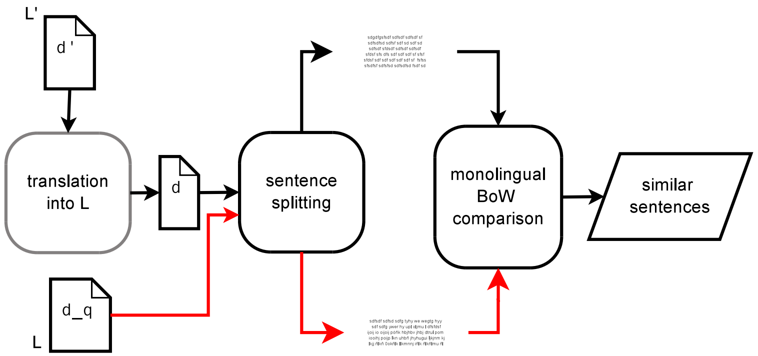
- Researchers are still forging the state of art in CL plagiarism detection
- The most of the methods are based on previously proposed models for CLIR

CL: Multilingualism

CL: Multilingualism



CL: Translation + Monolingual Analysis



The translation can be carried out on the basis of:

- Commercial MT systems (such as Google and Babelfish)
- Giza++, Moses, SRILM
  [Och and Ney, 2003, Koehn et al., 2007, Stolcke, 2002]
- Considering multiple translations per word [Muhr et al., 2010]

CL: Thesaurus based

EUROVOC Thesaurus-based

- Thesaurus catalogued manually
- Available in the 18 EU languages

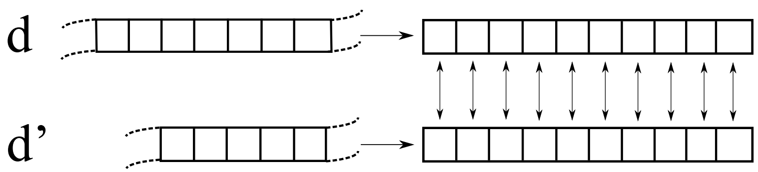
Example "transport of dangerous goods" lemmas

Lemma	Weight	Lemma	Weight
dangerous goods	33	radioactive material	19
by road	19	carriage	19
dangerous	18	plutonium	17
radioactive waste	15	nuclear fuel	15
shipment	15	adr	14
bind for	13	tank	13
receptacle	13	transport	13
pollute	12	nuclear waste	12

[Pouliquen et al., 2003]

CL: Thesaurus based

- $d \in L$ and $d' \in L'$ are mapped into a vector of thesaurus descriptor terms



$$sim(d, d') = cos(\theta_{\mathbf{d},\mathbf{d'}})$$

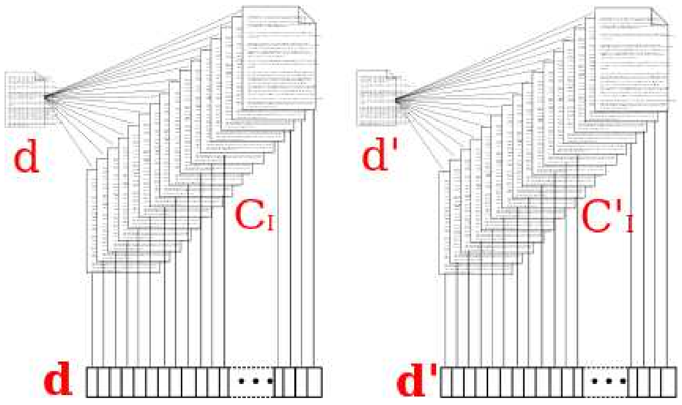
[Pouliquen et al., 2003]

CL: Explicit Semantic Analysis

- A significant comparable corpus C is required
- $d \in L$ ($d' \in L'$) is represented as a vector of relations to the index collection C_I (C'_I)
- The similarities are computed using a monolingual retrieval model such as the VSM
- Wikipedia is one of the biggest comparable corpora nowadays

[Potthast et al., 2008]

CL: Explicit Semantic Analysis



[Potthast et al., 2008]

CL: Alignment-based Similarity Analysis

- How likely is that d is a valid translation of d'?
- A two-step probabilistic translation and similarity analysis
- An adaptation of basic principles statistical MT

[Pinto et al., 2009]

CL: Alignment-based Similarity Analysis

Baye's rule for statistical Machine Translation:

$$p(d' \mid d_q) = \frac{p(d')\ p(d_q \mid d')}{p(d_q)}$$

- $p(d_q)$ does not depend on d' and is therefore neglected
- $p(d_q \mid d')$ is a translation model probability (statistical bilingual dictionary)
- $p(d')$ is the language model probability

[Brown et al., 1993]

CL: Alignment-based Similarity Analysis

$$p(d' \mid d_q) = p(d')\ p(d_q \mid d')$$

Two adaptations can be made:

- The adapted translation model is a non-probabilistic measure $w(d_q \mid d')$
- The language model is replaced by a length model $\varrho(d')$ that depends on document length

$$\varphi(d_q, d') = s(d' \mid d_q) = \varrho(d')\ w(d_q \mid d').$$

[Barrón-Cedeño et al., 2008, Pinto et al., 2009, Potthast et al., 2011]

CL: Alignment-based Similarity Analysis

The translation model depends on a bilingual dictionary (estimated by the IBM M1)

es	en	$p(es, en)$
certifica	certifies	0.420329
certifica	certify	0.164481
certifica	certified	0.109649
certifica	certifying	0.091375
certifica	hereby	0.054824
certifica	that	0.050577
certifica	has	0.035947
certifica	declare	0.018275
certifica	licence	0.018271

CL: Alignment-based Similarity Analysis

Translation model

$$p(d \mid d') = \prod_{x \in d} \sum_{y \in d'} p(x, y)$$

Adapted translation model (document level)

$$w(d \mid d') = \sum_{x \in d} \sum_{y \in d'} p(x, y)$$

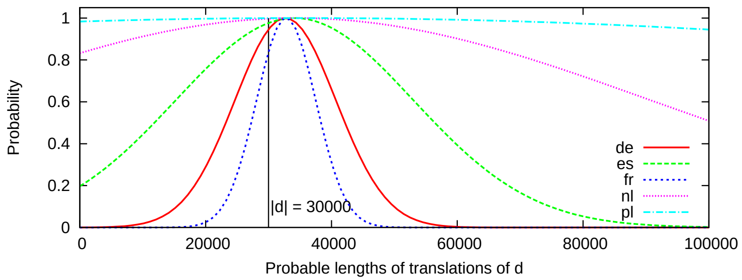
- $w(d \mid d')$ increases if valid translations (x, y) appear in the implied vocabularies.
- For a word x, with $p(x, y) = 0$ for all $y \in d'$, $w(d \mid d')$ is decreased by ε, in our case $\varepsilon = 0.1$.

CL: Alignment-based Similarity Analysis

Length Model

- It is expected that the length of the translation documents d and d' is closely related [Pouliquen et al., 2003]

$$\varrho(d') = e^{-0.5\left(\frac{\frac{|d'|}{|d|} - \mu}{\sigma}\right)^2}$$



[Potthast et al., 2011]

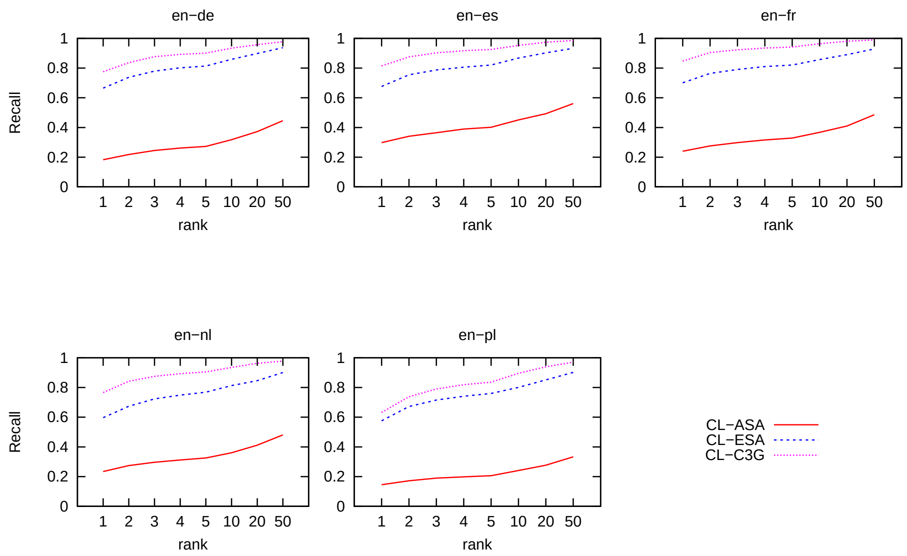
CL: Character n-grams

Character n-grams use to be common languages with syntactical similarities.

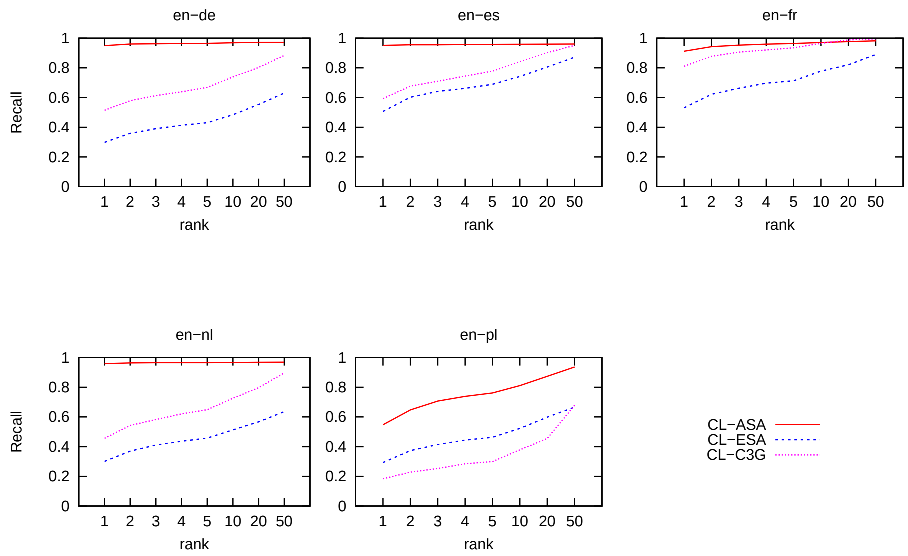
- $\Sigma = \{a, \ldots, z, 0, \ldots, 9\}$,
- $n = 3$
- $tfidf$-weighting
- Cosine similarity

[Mcnamee and Mayfield, 2004]

CL: Cross-Language Ranking (Wikipedia)



CL: Cross-language ranking (JRC-Acquis)



CL: And for less related languages?

The Party of European Socialists (PES) is a European political party comprising thirty-two socialist, social democratic and labour parties from each European Union member state and Norway.

El Partido Socialista Europeo (PSE) es un partido político pan-europeo cuyos miembros son de partidos socialdemócratas, socialistas y laboristas de estados miembros de la Unión Europea, así como de Noruega.

Europako Alderdi Sozialista Europar Batasuneko herrialdeetako eta Norvegiako hogeita hamahiru alderdi sozialista, sozialdemokrata eta laborista biltzen dituen alderdia da.

The corresponding articles contain around 2,000, 1,300, and only 100 words! [Wikipedia, 2010b]

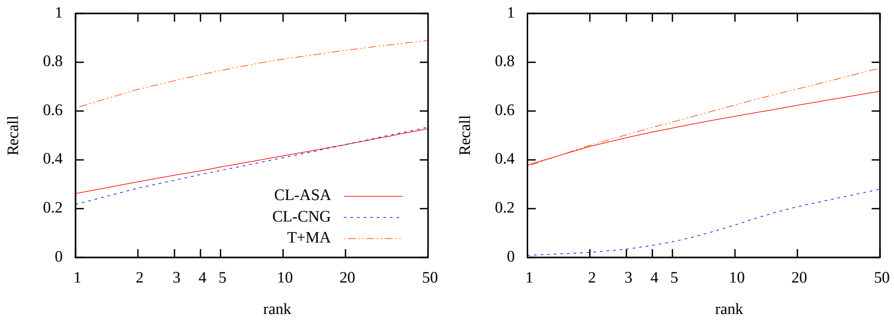
CL: Less Resourced Languages

Framework

- Two parallel corpora:
  - software a translation memory (en-eu)
  - consumer extracts from a multilingual magazine (es-eu)
- The entire corpus is a “big” document
- We perform sentence level similarity estimation

(corpora provided by Elhuyar Fundazioa and Consumer)

CL: Less Resourced Languages



(a) es-eu (b) en-eu

And these are not with Greek, Hindi, Chinese…!

[Barrón-Cedeño et al., 2010]

Outline

PAN

http://pan.webis.de

Potthast, et al. An Evaluation Framework for Plagiarism Detection. Coling 2010 (posters), pp. 997-1005.

[Potthast et al., 2009, Potthast et al., 2010a]

PAN-PC-09: Corpus of Synthetic Plagiarism

- Plagiarism implies an ethical issue
- Nobody would like to be included in a corpus containing plagiarism!
- Properly anonymising actual cases of plagiarism is a hard task
- Manual analysis should be necessary to define plagiarised-original text borders

PAN-PC-09: Corpus of Synthetic Plagiarism

Base texts: Texts from Project Gutenberg (http://www.gutenberg.org).

Restrictions: As the base text is free of copyright, the resulting corpus does not have distribution restrictions.

Cases generation: All the cases of text reuse are created automatically.

Proper citation: No cases of proper citation are included.

PAN-PC-09: Corpus of Synthetic Plagiarism

"A newly developed large-scale corpus of artificial plagiarism"

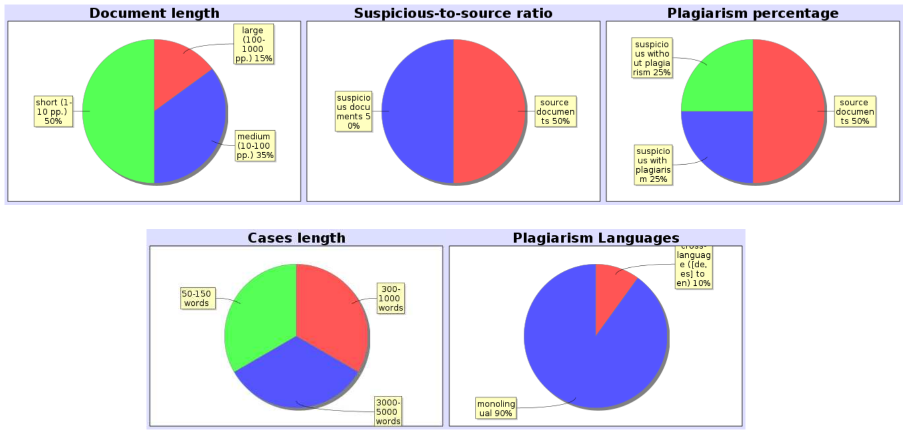
- 41 223 documents
- 94 202 artificial plagiarism cases
- It includes cases for intrinsic and external detection methods

http://www.webis.de/research/corpora

PAN-PC-09: Corpus Parameters

- Document length
- Suspicious-to-source ratio
- Plagiarism percentage
- Cases length
- Plagiarism language
- Cases obfuscation

PAN-PC-09: Corpus Parameters



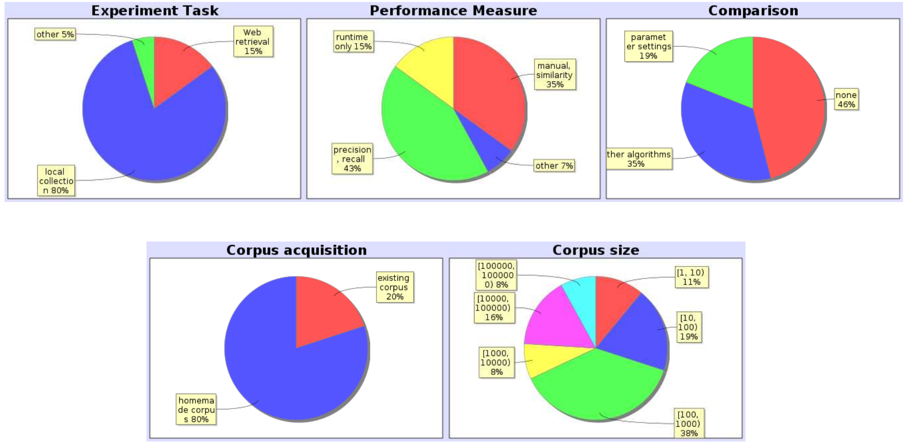
PAN-PC-09: Simulating Obfuscation

Cases Obfuscation

Paraphrasing, summarisation, etc. is simulated by…

- shuffling, removing, inserting short phrases
- replacing semantically related words
- POS preserving shuffling

PAN: How Researchers Evaluate Plag. Detection



[Potthast et al., 2010b]

PAN: How Researchers Evaluate Plag. Detection

- No standard evaluation measures have been previously defined
- Evaluations use to be incomparable and often not even reproducible
- How can we determine what model performs best?

PAN: Evaluation Measures

We are interested in evaluating three main aspects:

❶ plagiarised and —if available— source fragments are retrieved;

❷ original text fragments are not reported as plagiarised; and

❸ plagiarised fragments are not detected over and over again.

PAN: Evaluation Measures

Precision and Recall

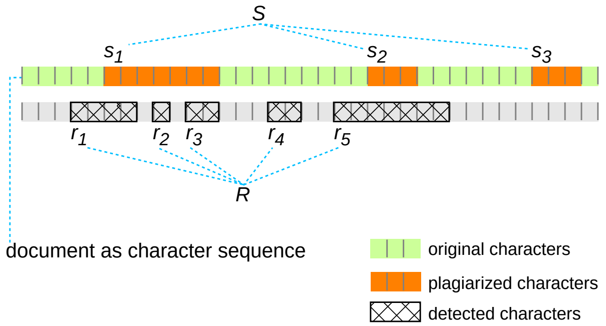
$$precision = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{retrieved\ documents\}|}$$

$$recall = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{relevant\ documents\}|}$$

F-measure

$$F = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

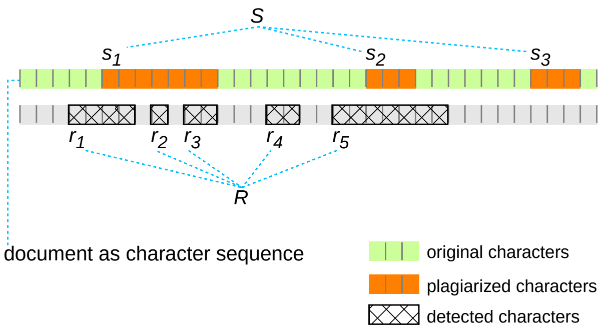
PAN: Evaluation Measures - P and R



$$rec_{PDA}(S,R) = \frac{1}{|S|}\sum_{s \in S} \frac{|s \sqcap \bigcup_{r \in R} r|}{|s|} \qquad prec_{PDA}(S,R) = \frac{1}{|R|}\sum_{r \in R} \frac{|r \sqcap \bigcup_{s \in S} s|}{|r|}$$

($\sqcap$ computes the positionally overlapping characters)

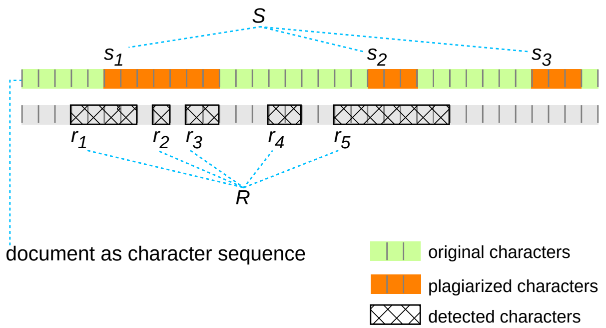
PAN: Evaluation Measures - Granularity



$$gran_{PDA}(S,R) = \frac{1}{|S_R|}\sum_{s \in S_R} |C_s| \quad \in [1, |R|]$$

$C_s = \{r \mid r \in R \ \wedge \ s \cap r \neq \emptyset\}$
$S_R = \{s \mid s \in S \ \wedge \ \exists r \in R: \ s \cap r \neq \emptyset\}$

PAN: Evaluation Measures - plagdet



$$plagdet_{PDA}(S,R) = \frac{F}{\log_2(1 + gran_{PDA})}$$

PAN: 1st International Competition - Game Rules

Eligibility The contest was open to any party planning to attend the PAN competition. No feedback at the time of submission was provided.

Integrity The exploitation of potential flaws in the competition corpus to gain advantages was prohibited.

Text resources No other text than the one provided in the corpus could be used.

Winner Selection One winner of the “External Plagiarism Detection” task, one winner of the “Intrinsic Plagiarism Detection” task, and one overall winner were proclaimed.

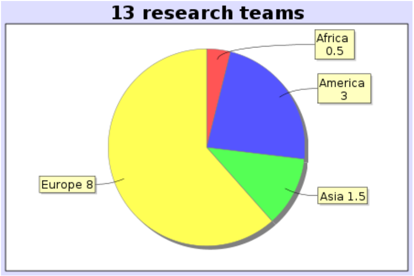
Award The overall winner was awarded a prise, sponsored by Yahoo! Research.

PAN: 1st International Competition - Chronology

March 2009 Participants were provided with the developing section of the corpus (with annotated cases).

May 2009 Test corpus provided (without any annotation).

June 2009 Participants submitted their detections to be evaluated.



PAN: 1st International Competition, Overview

Intrinsic Approaches (4 teams)

Participant	Analysed features
Stamatatos	character n-grams
Zechner, Muhr, Kern, Granitzer	word freq. class + text frequencies
Seaward, Matwin	Kolmogorov complexity measures

External Approaches (10 teams)

Participant	Comparison units
Grozea, Gehl, Popescu	character n-grams
Kasprzak, Brandejs, Kripac	word n-grams
Basile, Benedetto, Caglioti, Degli Esposti	length n-grams

http://www.webis.de/research/workshopseries/pan-09/competition.html

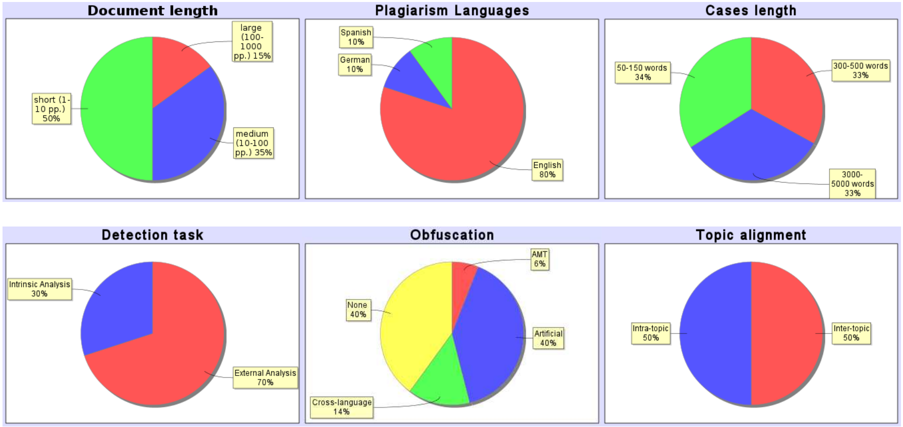
http://ceur-ws.org/Vol-502

PAN-PC-10 Corpus

- $27,073$ documents (obtained from 22 874 books from the Project Gutenberg2)
- $68,558$ plagiarism cases (about 0-10 cases per document)

www.webis.de/research/corpora/pan-pc-10

PAN-PC-10 Corpus Parameters



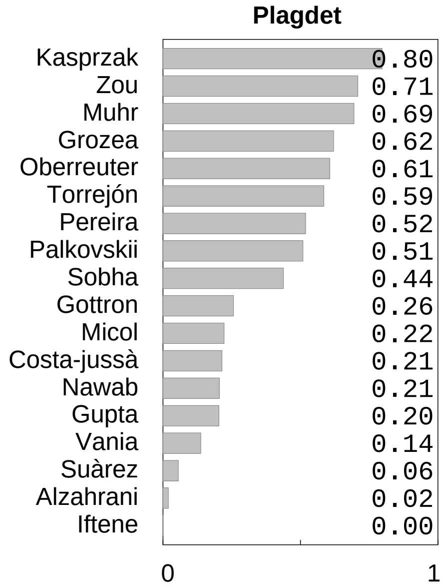
PAN: 2nd International Competition

March 2010 Participants were provided with the developing section of the corpus (PAN-PC-09)

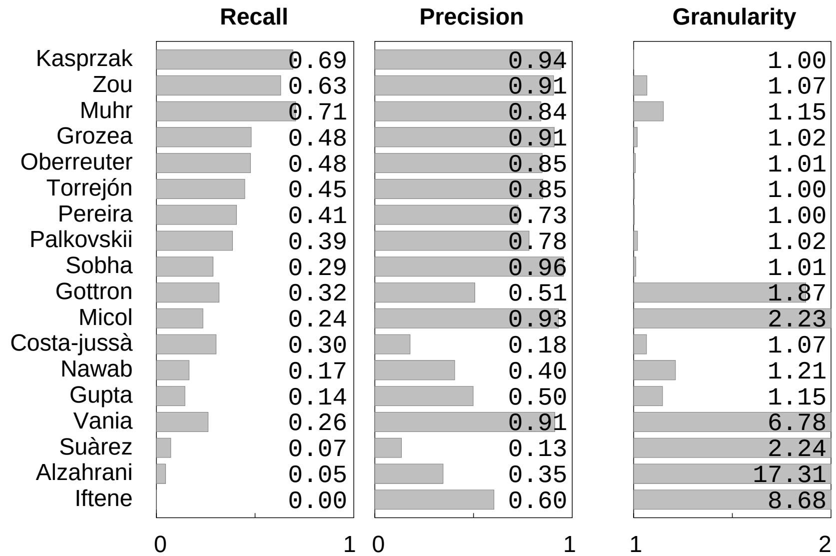
May 2010 Test corpus provided (brand new)

June 2010 Participants submitted their detections to be evaluated.

PAN: 2nd International Competition Results



PAN: 2nd International Competition Results



Outline

Introduction

Basic Concepts

Intrinsic Plagiarism Detection

External Plagiarism Detection

Cross-Language Plagiarism Detection

Plagiarism Detection Competition

Not Only Plain Text, Not only Plagiarism

Start Point

Cutting the Edge

Not Only…: Software Plagiarism

A program that has been produced from another with a small number of routine transformations.

Student plagiarism reasons:

1990's
- large undergraduate classes,
- introduction of personal computers,
- computer networks,
- easy-to-use screen editors

Today
- Internet

[Parker and Hamblen, 1989]

Not Only…: Software Plagiarism

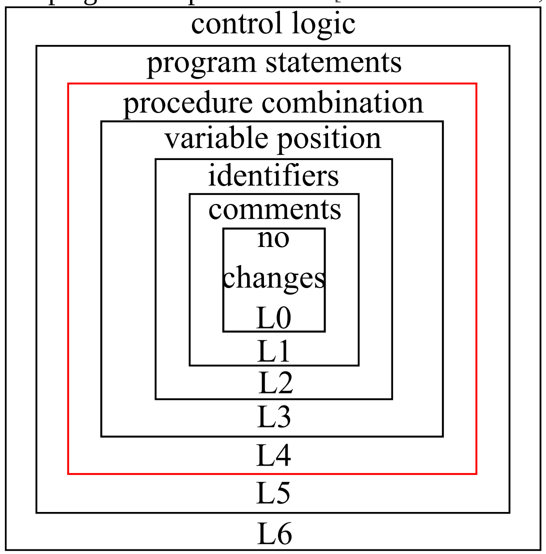
Techniques to disguise plagiarism

Operation	Example		
changing comments	//	→	/* */
changing formatting	Indentation		
changing identifiers	int x;	→	int y;
changing operands order	x<y	→	y≥x
changing data types	float x;	→	double x;
replacing expressions	printf…	→	echo…
adding redundant statements			
changing the order of statements	x=5; y=2*x;	→	y=10; x=y/2
changing the structure of iterations	for if	→	if for
changing the structure of selections	if…elif…else	→	switch
replacing function calls for functions			
combining original/copied sections			

[Whale, 1986]

Not Only…: Software Plagiarism

Program plagiarism spectrum [Faidhi and Robinson, 1987]



Not Only…: Plagiarism

Feature	dependent	independent
characters per line		■
comment lines		■
indented lines	■	
blank lines		■
avg. function length		■
reserved words	■	
avg. identifier length		■
avg. space per line (%)		■
total operands		■
total operators		■
conditional statement (%)	■	
repetitive statement (%)	■	
multiple statement lines		■

Some (statistical) features

[Parker and Hamblen, 1989]

Not Only…: Software Plagiarism

YAP

- Comments and string-constants are removed.
- Upper-case letters are translated to lower-case
- If possible, the functions/procedures are expanded in calling order.
- Tokens not in the lexicon for the language are removed.
- Greedy string comparison

http://luggage.bcs.uwa.edu.au/~michaelw/YAP.html

[Parker and Hamblen, 1989]

Not Only…: Source Code Analysis Tools

MOSS ✓

- Based on fingerprinting
- http://theory.stanford.edu/~aiken/moss/

JPLAG ✓

- Based on Greedy String Tiling
- www.ipd.uni-karlsruhe.de/jplag

Cogger ·

- Case based reasoning (the problem of finding similarity in programs is made analogous to the problem of case retrieval)

Not Only…: CL Source Code Analysis

Cross-language plagiarism makes sense in programming languages?

- A person could “copy” a program from a language into another one
- Can we detect if a program is the implementation of some algorithm pseudo-code? (consider that often “pseudo-code” is in fact Python or some simplified programming language)
- Maybe a programmer is fired and we want to check if he already coded the algorithm we asked…

However, most methods simply apply tokenisation and string matching comparison

Not Only…: CL Source Code Analysis

```
if (score < 60) {
   comment = "This is terrible";
}
else {
   comment = "Not so bad";
}
```

```
if score < 60:
   comment = "This is terrible"
elif score == 60:
   comment = "This is bad"
else:
   comment = "Not so bad"
```

```
if ($score < 60) {
  $comment = "This is terrible";
}
elsif ($score == 60) {
   $comment = "This is bad";
}
else{
   $comment = "Not so bad";
}
```

```
if (score < 60) {
   comment = "This is terrible";
}
else {
   comment = "Not so bad";
}
```

```
if ($score < 60) {
  $comment = "This is terrible";
}
elsif ($score == 60) {
   $comment = "This is bad";
}
else{
   $comment = "Not so bad";
}
```

```
If score < 60 Then
   comment = "This is terrible"
Elseif score == 60 Then
   comment = "This is bad"
else
   comment = "Not so bad"
End If
```

```
if score < 60
   comment = "This is terrible"
elsif score == 60
   comment = "This is bad"
else
   comment = "Not so bad"
end
```

```
if [$score  <  60]; then
   $comment = "This is terrible"
else
   $comment = "Not so bad"
fi
```


Not Only…: CL Source Code Analysis

X-plag

The only method for CL programming plagiarism detection (we are aware of)

- Instead of comparing the source codes, it compares "intermediate code"

  .NET Visual{C#, Basic.NET, J#, C++.NET}

  GCC C, C++, Java, Fortran, Objective C

RTL: Register Transfer Language, a common intermediate code (GCC)

[Arwin and TahaGhoghi, 2006]

Not Only…: X-plag

Detection Process

- Intermediate code generation
- Filtering process (just a set of keywords is considered relevant)
- Comparison based on n-grams!

Not Only…: Actual CL Analysis in Source Code?

```
... 60) {
   comment = "This is terrib...
}
else {
   comment = "Not so bad";
```

```
if score < 60:
   comment = "This is terrible"
elif score == 60:
   comment = "This is bad"
else:
   comment = "Not so bad"
```

```
if ($score < 60) {
  $comment = "This is terrible";
}
elsif ($score == 60) {
   $comment = "This is bad";
}
else{
   $comment = "Not so bad";
}
```

```
... 60) {
   comment = "This is terri...
}
else {
   comment = "Not so bad";
```

```
if ($score < 60) {
  $comment = "This is terrible";
}
elsif ($score == 60) {
   $comment = "This is bad";
}
else{
   $comment = "Not so bad";
}
```

```
If score < 60 Then
   comment = "This is terrible"
Elseif score == 60 Then
   comment = "This is bad"
else
   comment = "Not so bad"
End If
```

```
if score < 60
   comment = "This is terrible"
elsif score == 60
   comment = "This is bad"
else
   comment = "Not so bad"
end
```

```
if [$score  <  60]; then
   $comment = "This is terrible"
else
   $comment = "Not so bad"
fi
```

?

Not Only…: Actual CL Analysis in Source Code?

1. Is there a length factor between programming languages?
   - C and Java lengths are closed…
   - Python is shorter…
2. Is it possible to learn a bilingual dictionary of programming languages?
   - print printf 0.9; print echo 0.05…
3. Could we use a method such as CL-ESA?
4. BTW: What about plagiarised methods/functions? (not entire programs)

Not Only…: Wikipedia Revisions

- Different Wikipedias have different behaviour
- Not plagiarism, but collaborative authorship

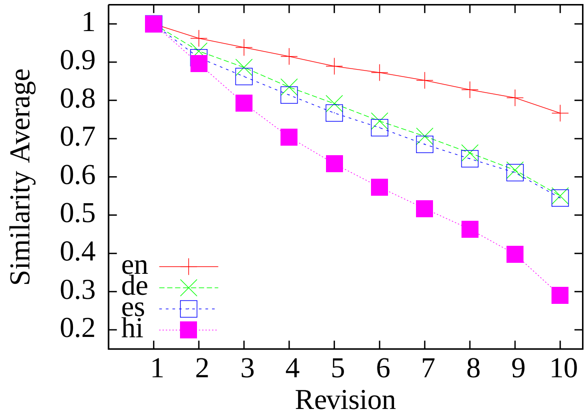
Corpus

- Wikipedia articles in: English, German, Spanish, and Hindi
- The 500 most “popular” articles were considered
- 10 revisions considered per article

Not Only…: Wikipedia Revisions Corpus

Lan	$\|D_q\|$	$\|D\|$	$\|d_{avg}\|_t$	$\|d_{avg}\|$	$\|D\|_t$
	Before stopwords elimination				
de	500	5,000	1,812	5,229	261,370
en	500	5,000	2,243	8,552	183,414
hi	500	5,000	302	672	78,673
es	500	5,000	1,216	4,116	133,595
	After stopwords elimination				
de	500	5,000	1,707	3,474	261,146
en	500	5,000	2,149	6,008	183,288
hi	500	5,000	270	495	78,577
es	500	5,000	1,142	2,415	133,339

Not Only…: Wikipedia Revisions Evolution



Not Only…: Wikipedia Revisions - Experiments

1. Document level
   - For each document $d_q \in D_q$ the documents in D are ranked with respect to $sim(d, d_q)$, generating r_q
   - We expect that d_q is co-derived from the documents on top of r_q.
2. Section level
   - The sections in the top-50 of r_q compose the set D' of co-derivative candidate sections.
   - D'_q is composed of the sections in $d_q \in D_q$.
   - For each section $d'_q \in D'_q$ the sections in D' are ranked with respect to their similarity $sim(d', d'_q)$.
   - It is expected that those sections in the top of r'_q are actual co-derivatives of d'_q.

Not Only…: Wikipedia Revisions - Metrics

- $P@10$ and $R@m$ by considering $m = \{10, 20, 50\}$
- $P@10 = R@10$

Highest False Match and Separation

- Estimate the distance of the correctly and incorrectly retrieved documents in r_q
- The calculation is possible only if $R@50 = 1.0$

[Hoad and Zobel, 2003]

Not Only…: Wikipedia Revisions - Metrics

Highest False Match and Separation

$$HFM = \frac{100 \cdot sim(d^-, d_q)}{s^*}$$

- s^* is the maximum similarity value
- d^- is the highest ranked document which is not relevant concerning d_q

$$sep = \frac{100 \cdot (sim(d^+, d_q) - sim(d^-, d_q))}{s^*}$$

- d^+ is the lowest ranked document which is relevant concerning d_q
- $LTM = 100 \cdot sim(d^+, d_q)/s^*$ is the Lowest True Match

- $sep > 0 \Rightarrow$ the highest rated documents in r_q are all relevant
- $sep < 0 \Rightarrow$ other documents were ranked before those relevant

[Hoad and Zobel, 2003]

Not Only…: Wikipedia Revisions - Results

J - Jaccard
C - Cosine
K - Kullback Leibler
M - M. Translation
O - Okapi BM25
P - W. chunk. overlap
W - Winnowing
S - Spex

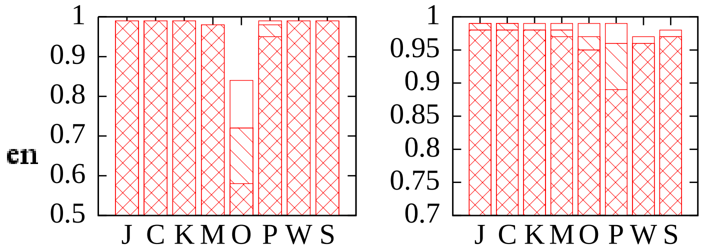
R@50
R@20
R@10

s_d sep. for documents
s_s sep. for sections
H_d HFM for documents
H_s HFM for sections

Document **Section**

HFM, sep

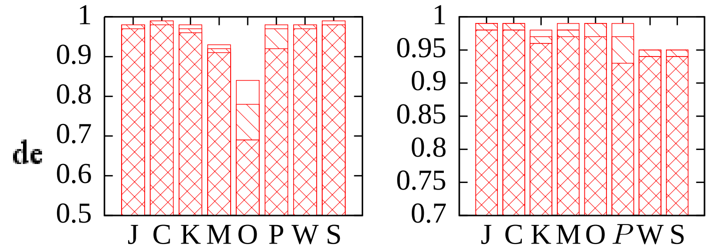
Not Only…: Wikipedia Revisions - English



	J	C	K	M	O	P	W	S
S_d	52.6	39.9	25.1	42.3	5.7	21.3	37.6	61.5
S_s	64.7	44.7	23.5	59.1	49.3	27.8	52.8	71.4
H_d	23.4	54.6	68.3	37.9	71.2	64.3	1.7	3.7
H_s	15.7	47.0	65.2	26.1	40.4	60.7	1.9	12.8

Worst
Better

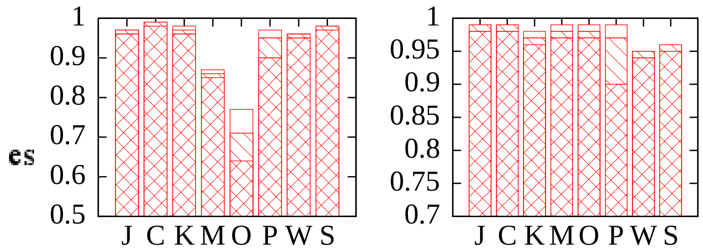
Not Only…: Wikipedia Revisions - German



	J	C	K	M	O	P	W	S
S_d	42.1	40.4	20.0	40.5	14.9	10.9	25.8	44.9
S_s	63.6	50.8	25.7	66.6	55.7	36.1	56.7	68.3
H_d	14.6	42.1	65.1	13.7	58.2	64.6	0.60	1.40
H_s	12.1	36.5	66.1	66.7	30.8	51.3	1.08	1.90

Worst
Better

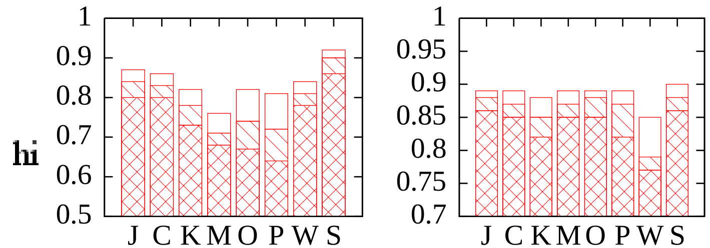
Not Only…: Wikipedia Revisions - Spanish



	J	C	K	M	O	P	W	S
S_d	39.7	39.6	19.8	33.3	10.0	17.2	32.4	48.1
S_s	66.7	50.9	26.4	63.7	53.9	32.1	65.6	75.4
H_d	15.2	40.1	62.3	36.1	56.9	54.3	0.73	1.8
H_s	13.0	38.8	67.1	20.7	34.6	56.5	0.77	1.7

Worst
Better

Not Only…: Wikipedia Revisions - Hindi

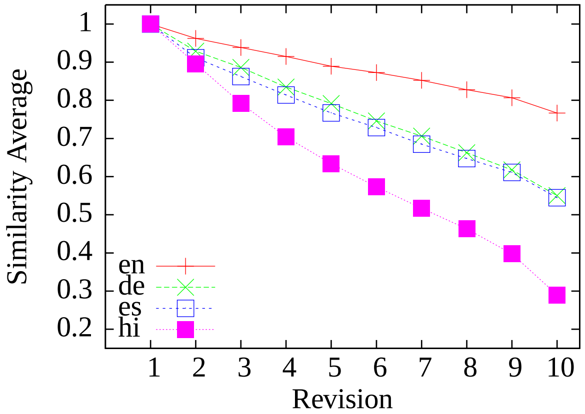


	J	C	K	M	O	P	W	S
S_d	18.7	19.0	7.4	17.2	9.8	8.9	11.9	29.7
S_s	52.5	41.1	23.5	49.9	44.3	34.6	57.2	66.9
H_d	18.8	41.0	65.2	29.9	49.7	62.5	8.3	14.7
H_s	20.9	43.2	66.4	29.5	38.4	50.3	4.2	8.0

Worst
Better

Not Only…: Wikipedia Revisions - Observations

Is Hindi a more difficult language to work with?



Not Only…: Wikipedia Revisions - Observations

Document Level	Section Level
• Best: Fingerprinting models	• Best: Jaccard, Cosine, ∼IBM1

If all the relevant documents are in the top-50

• Best: Cosine and KL	• Best: Okapi, Jaccard and Cosine

Outline

Start Point: Try with “Small” Corpora

1. METER http://www.dcs.shef.ac.uk/nlp/meter/
   - Advantages
     - Small amount of documents
     - verbatim/modified copy and new fragments identified
     - Real cases of journalistic text reuse manually analysed
   - Disadvantage
     - No low level annotation (fragments)
2. Co-derivatives http://www.dsic.upv.es/grupos/nle/
   - Advantages
     - Small amount of documents
     - Documents relations identified
     - Includes different languages (even Hindi)
   - Disadvantages
     - No low level annotation (fragments)
     - Wikipedia revisions are far from realistic text reuse
3. CLiPA http://www.dsic.upv.es/grupos/nle/
   - Advantages
     - Contains cross-language text reuse cases
     - Created with humans and MT systems
   - Disadvantage
     - Extremely small (toy corpus)

Start Point: PAN trends

- Study the proceedings of the first two competitions:

http://pan.webis.de

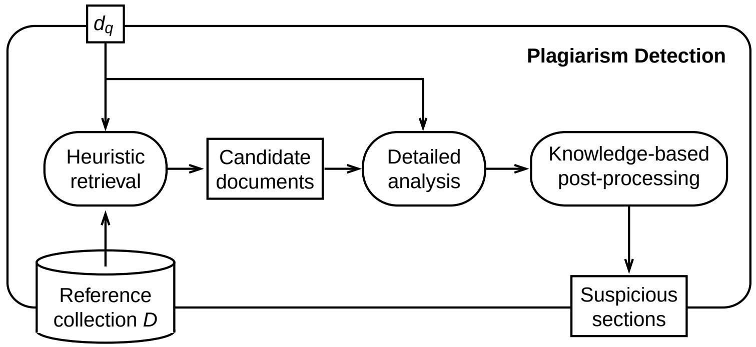
- Prove your own models over the PAN-PC-09 and PAN-PC-10
- Focus on developing good models instead of winning a competition!

Start Point: Video

Et Plagieringseventyr. Universitetet i Bergen, http://sokogskriv.no/english/
http://www.youtube.com/watch?v=Mwbw9KF-ACY

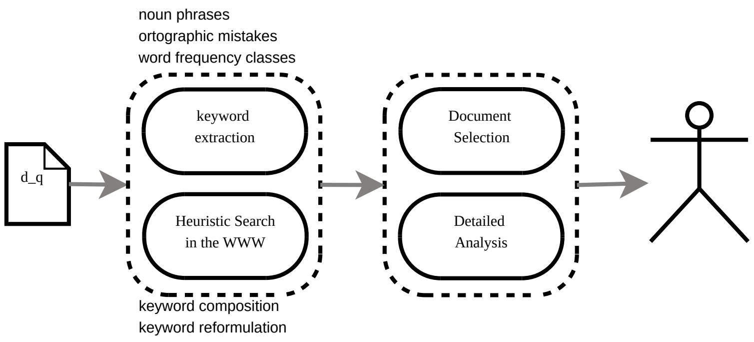
Outline

Edge: Plagiarism Detection Process



[Stein et al., 2007]

Edge: Plagiarism Detection Process (revisited)



(adapted from Stein's keynote speech at SEPLN 2010)

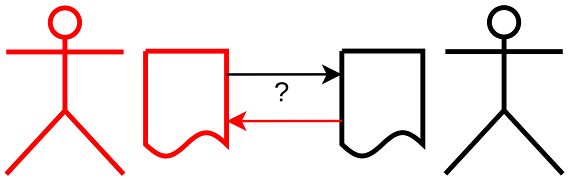
Edge: Improving Models

- Improve document access
- Improve processing time

(adapted from Stein's keynote speech at SEPLN 2010)

- Improve Cross-Language models
  [Barrón-Cedeño et al., 2008, Barrón-Cedeño et al., 2010, Ceska et al., 2008, Lee et al., 2008]
- Create better intrinsic analysis models
  [Meyer zu Eißen and Stein, 2006, Stamatatos, 2009]

Edge: Who's the Thief?



- Perform intrinsic analysis over the two documents. That document with variations between the alleged reused fragment is the thief
- Use an adaptation of Encoplot

[Grozea and Popescu, 2010]

Edge: Identifying Proper Citation

- People reuse text from others (this is a fact!)
- However, sometimes they include proper citation

"All of the books in the world contain no more information than is broadcast as video in a single large American city in a single year. Not all bits have equal value." Carl Sagan

As Groucho Marx said in his book Groucho and Me (1959), "no one is completely unhappy at the failure of his best friend".

Post processing Divide cases of reuse with proper citation from actual plagiarism

Edge: Wikipedia Multilingual Reuse

- In Wikipedia articles in different topics are available in hundreds of languages.
- English Wikipedia is the most developed: $\sim$3.4M articles (only comparable to the sum of German, French, Polish, and Italian Wikipedias altogether)
- It has been referred as one of the hugest comparable corpus at hand [Mohammadi and GhasemAghaee, 2010, Potthast et al., 2008].

Edge: Wikipedia Multilingual Reuse

Comparable Corpora

- it contains the same proportions of texts of the same genres, same domains and in a range of different languages; and
- such texts are sampled on the same period.

Parallel $\longrightarrow$ Comparable $\longrightarrow$ Non Parallel

- parallel corpus: sentence aligned corpus containing bilingual translations of the same document;
- noisy parallel corpus: includes aligned and non-aligned sentences;
- comparable corpus: collection which does not contain aligned sentences, but which is about the same topic;
- non parallel corpus: collection containing disparate bilingual documents, which may or may not be on the same topic.

[McEnery and Xiao, 2007, Fung and Cheung, 2004]

Edge: Wikipedia Multilingual Reuse

- Reuse of text across related articles
- Reuse of text outside of Wikipedia (related to [Bendersky and Croft, 2009])
- Cross-language text reuse

Edge: Creating More Resources

- Create more and better corpora
- Increase the amount of cross-language cases
- Create (simulated) human made cases

Activity 3: Creating Cases of Cross-Language Plagiarism Detection

Images Sources

1 http://www.baltimoreegypt.org
2 http://clatterymachinery.wordpress.com
3 http://www.berkshirehistory.com/bios/apope.html
4 http://www.hoasm.org/IVM/Jonson.html
5 http://fcom.us.es/blogs/vazquezmedel/tag/samuel-johnson/
6 http://toosweet4rocknroll.wordpress.com
7 [Basile et al., 2009]
8 http://www.dailymail.co.uk
9 http://www.levante-emv.com/
10 http://www.lavanguardia.es/
11–16 http://www.wikimedia.org

Thank you!

Alberto Barrón Cedeño and Paolo Rosso

[lbarron,prosso]@dsic.upv.es

http://www.dsic.upv.es/grupos/nle

References I

Arwin, C. and TahaGhoghi, S. (2006).
Plagiarism Detection across Programming Languages.
In Proceedings of the Australasian Computer Science Conference (ACSC 2006), Tasmania, Australia.

Barrón Cedeño, A. (2008).
Detección automática de plagio en texto.
Master's thesis, Universidad Politécnica de Valencia, Valencia, España.
Supervisor: Paolo Rosso.

Barrón-Cedeño, A. and Rosso, P. (2009).
On Automatic Plagiarism Detection based on n-grams Comparison.
Advances in Information Retrieval. Proceedings of the 31st European Conference on IR Research, LNCS (5478):696–700.

Barrón-Cedeño, A., Rosso, P., Agirre, E., and Labaka, G. (2010).
Plagiarism Detection across Distant Language Pairs.
In Huang, C.-R. and Jurafsky, D., editors, Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010). Coling 2010 Organizing Committee.

Barrón-Cedeño, A., Rosso, P., Pinto, D., and Juan, A. (2008).
On Cross-lingual Plagiarism Analysis Using a Statistical Model.
In Stein, B., Stamatatos, E., and Koppel, M., editors, ECAI 2008 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 2008), pages 9–13. CEUR-WS.org.

Basile, C., Benedetto, D., Caglioti, G., and Degli Esposti, M. (2009).
A Plagiarism Detection Procedure in Three Steps: Selection, Matches and Squares.
In [Stein et al., 2009], pages 19–23.

Bendersky, M. and Croft, W. (2009).
Finding Text Reuse on the Web.
In Proceedings of the Second ACM International Conference on Web Search and Data Mining, pages 262–271. ACM.

References II

Berger, A. and Lafferty, J. (1999).
Information Retrieval as Statistical Translation.
In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 222–229. ACM.

Bernstein, Y. and Zobel, J. (2004).
A Scalable System for Identifying Co-Derivative Documents.
In Proceedings of the Symposium on String Processing and Information Retrieval, pages 55–67. Springer.

Bigi, B. (2003).
Using Kullback-Leibler Distance for Text Categorization.
In Proceedings of the 25th ECIR'03, Springer-Verlag, volume LNCS (2633) Advances in Information Retrieval, pages 305–319, Pisa, Italy.

Braschler, M. and Harman, D., editors (2010).
Notebook Papers of CLEF 2010 LABs and Workshops, Padua, Italy.

Brin, S., Davis, J., and Garcia-Molina, H. (1995).
Copy Detection Mechanisms for Digital Documents.
In Carey, M. and Schneier, D., editors, Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data, pages 398–409. ACM Press.

Broder, A. (1997).
On the Resemblance and Containment of Documents.
In Compression and Complexity of Sequences (SEQUENCES'97), pages 21–29. IEEE Computer Society.

Brown, P., Della Pietra, S., Della Pietra, V., and Mercer, R. (1993).
The Mathematics of Statistical Machine Translation: Parameter Estimation.
Computational Linguistics, 19(2):263–311.

References III

Ceska, Z., Toman, M., and Jezek, K. (2008).
Multilingual Plagiarism Detection.
In Proceedings of the 13th International Conference on Artificial Intelligence, pages 83–92. Springer Verlag Berlin Heidelberg.

Clough, P. and Gaizauskas, R. (2009).
Corpora and Text Re-Use.
In Lüdeling, A., Kytö, M., and McEnery, T., editors, Handbook of Corpus Linguistics, Handbooks of Linguistics and Communication Science, pages 1249—1271. Mouton de Gruyter.

Clough, P., Gaizauskas, R., and Piao, S. (2002).
Building and Annotating a Corpus for the Study of Journalistic Text Reuse.
In Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002), volume V, pages 1678–1691.

Comas, R. and Sureda, J., editors (2008).
Academic cyberplagiarism, volume 10 of Digithum.
Universitat Oberta de Catalunya.

Faidhi, J. and Robinson, S. (1987).
An empirical approach for detecting program similarity and plagiarism within a university programming environment.
Comput. Educ., 11(1).

Fung, P. and Cheung, P. (2004).
Mining very non-parallel corpora: Parallel sentence and lexicon extraction vie bootstrapping and EM.
In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pages 57—63.

Goldstein, J., Mittal, V., Carbonell, J., and Kantrowitz, M. (2000).
Multi-Document Summarization By Sentence Extraction.
In NAACL-ANLP 2000 Workshop on Automatic Summarization, pages 40–48, Seattle, WA. Association for Computational Linguistics.

References IV

Grozea, C., Gehl, C., and Popescu, M. (2009).
ENCOPLOT: Pairwise Sequence Matching in Linear Time Applied to Plagiarism Detection.
In [Stein et al., 2009], pages 10–18.

Grozea, C. and Popescu, M. (2010).
Who's the Thief? Automatic Detection of the Direction of Plagiarism.
Computational Linguistics and Intelligent Text Processing, 10th International Conference, LNCS (6008):700–710.

Hoad, T. and Zobel, J. (2003).
Methods for Identifying Versioned and Plagiarized Documents.
Journal of the American Society for Information Science and Technology, 54(3):203–215.

IEEE (2008).
A plagiarism FAQ.
http://www.ieee.org/web/publications/rights/plagiarism_FAQ.htm.
[Online; accessed 3-March-2010].

Irribarne, R. and Retondo, H. (1981).
Plagio de obras literarias. Inícitis Civiles y Penales en Derecho de Autor.
IIDA, Buenos Aires, Argentina.

Jaccard, P. (1901).
Étude comparative de la distribution florale dans une portion des Alpes et des Jura.
Bulletin del la Société Vaudoise des Sciences Naturelles, 37:547–579.

Kang, N., Gelbukh, A., and Han, S. (2006).
PPChecker: Plagiarism Pattern Checker in Document Copy Detection.
In Sojka, P., Kopeček, I., and Pala, K., editors, Proceedings of the Text, Speech and Dialogue, 10th International Conference (TSD 2006), volume LNCS (LNAI) (4188), pages 661–667. Springer-Verlag.

References V

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007).
Moses: Open Source Toolkit for Statistical Machine Translation.
In Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session.

Kulathuramaiyer, N. and Maurer, H. (2007).
Coping With the Copy-Paste-Syndrome.
In E-Learn 2007, pages 1072—1079, Quebec, CA.

Kullback, S. and Leibler, R. (1951).
On Information and Sufficiency.
Annals of Mathematical Statistics, 22(1):79–86.

Lee, C., Wu, C., and Yang, H. (2008).
A Platform Framework for Cross-lingual Text Relatedness Evaluation and Plagiarism Detection.
In Proceedings of the 3rd International Conference on Innovative Computing Information (ICICIC'08). IEEE Computer Society.

Lynch, J. (2006).
The Perfectly Acceptable Practice of Literary Theft: Plagiarism, Copyright, and the Eighteenth Century.
Colonial Williamsburg.

Maurer, H., Kappe, F., and Zaka, B. (2006).
Plagiarism - A Survey.
Journal of Universal Computer Science, 12(8):1050–1084.

McEnery, A. and Xiao, Z. (2007).
Parallel and Comparable Corpora: What Are They Up To?
In Rogers, M. and Anderman, G., editors, Incorporating Corpora. The Linguist and the Translator, pages 18–31. Clevedon.

References VI

Mcnamee, P. and Mayfield, J. (2004).
Character N-Gram Tokenization for European Language Text Retrieval.
Information Retrieval, 7(1-2):73–97.

Metzler, D., Bernstein, Y., Croft, W. B., Moffat, A., and Zobel, J. (2005).
Similarity Measures for Tracking Information Flow.
In Chowdhury, Fuhr, Ronthaler, Schek, and Teiken, editors, Proceedings of the 14th ACM International Conference on Information and Knowledge Management, pages 517–524, Bremen, Germany. ACM Press.

Meyer zu Eißen, S. and Stein, B. (2006).
Intrinsic Plagiarism Detection.
Advances in Information Retrieval: Proceedings of the 28th European Conference on IR Research (ECIR 2006), LNCS (3936):565–569.

Mohammadi, M. and GhasemAghaee, N. (2010).
Building Bilingual Parallel Corpora based on Wikipedia.
In Second International Conference on Computer Engineering and Applications, volume 2, pages 264–268.

Muhr, M., Kern, R., Zechner, M., and Granitzer, M. (2010).
External and Intrinsic Plagiarism Detection using a Cross-Lingual Retrieval and Segmentation System.
In [Braschler and Harman, 2010].

Och, F. and Ney, H. (2003).
A Systematic Comparison of Various Statistical Alignment Models.
Computational Linguistics, 29(1):19–51.
See also http://www.fjoch.com/GIZA++.html.

Parker, A. and Hamblen, J. (1989).
Computer Algorithms for Plagiarism Detection.
IEEE Transactions on Education, 32(2):94–99.

References VII

Pinto, D., Civera, J., Barrón-Cedeño, A., Juan, A., and Rosso, P. (2009).
A Statistical Approach to Crosslingual Natural Language Tasks.
Journal of Algorithms, 64(1):51–60.

Potthast, M., Barrón-Cedeño, A., Eiselt, A., Stein, B., and Rosso, P. (2010a).
Overview of the 2nd International Competition on Plagiarism Detection.
In [Braschler and Harman, 2010].

Potthast, M., Barrón-Cedeño, A., Stein, B., and Rosso, P. (2011).
Cross-Language Plagiarism Detection.
Language Resources and Evaluation, Special Issue on Plagiarism and Authorship Analysis.

Potthast, M., Stein, B., and Anderka, M. (2008).
A Wikipedia-Based Multilingual Retrieval Model.
In Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., and White, R., editors, 30th European Conference on IR Research, ECIR 2008, volume 4956 LNCS of Lecture Notes in Computer Science, pages 522–530, Berlin Heidelberg New York. Springer.

Potthast, M., Stein, B., Barrón-Cedeño, A., and Rosso, P. (2010b).
An Evaluation Framework for Plagiarism Detection.
In Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), Beijing, China.

Potthast, M., Stein, B., Eiselt, A., Barrón-Cedeño, A., and Rosso, P. (2009).
Overview of the 1st International Competition on Plagiarism Detection.
In [Stein et al., 2009], pages 1–9.

Pouliquen, B., Steinberger, R., and Ignat, C. (2003).
Automatic Identification of Document Translations in Large Multilingual Document Collections.
In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-2003), pages 401–408.

References VIII

Rodríguez Torrejon, D. and Martín Ramos, J. (2010a).
CoReMo System (Contextual Reference Monotony)n.
In [Braschler and Harman, 2010].

Rodríguez Torrejon, D. and Martín Ramos, J. (2010b).
Detección de plagio en documentos. Sistema externo monolingüe de altas prestaciones basado en n-gramas contextuales.
Procesamiento del Lenguaje Natural, 45:49–57.

Schleimer, S., Wilkerson, D., and Aiken, A. (2003).
Winnowing: Local Algorithms for Document Fingerprinting.
In Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data, New York, NY. ACM.

Shivakumar, N. and García-Molina, H. (1995).
SCAM: A Copy Detection Mechanism for Digital Documents.
In Proceedings of the 2nd Annual Conference on the Theory and Practice of Digital Libraries.

Spärck Jones, K., Walker, S., and Robertson, S. (2000).
A probabilistic model of information retrieval: development and comparative experiments.
Inf. Process. Manage., 36(6):779–840.

Stamatatos, E. (2009).
Intrinsic Plagiarism Detection Using Character n-gram Profiles.
In [Stein et al., 2009], pages 38–46.

Stein, B. (2005).
Fuzzy-Fingerprints for Text-Based Information Retrieval.
In Tochtermann, K. and Maurer, H., editors, Proceedings of the 5th International Conference on Knowledge Management (I-KNOW 2005), Journal of Universal Computer Science, pages 572–579. Know-Center.

References IX

Stein, B., Meyer zu Eissen, S., and Potthast, M. (2007).
Strategies for Retrieving Plagiarized Documents.
In Clarke, C., Fuhr, N., Kando, N., Kraaij, W., and de Vries, A., editors, Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 825–826, Amsterdam, The Netherlands. ACM.

Stein, B., Rosso, P., Stamatatos, E., Koppel, M., and Agirre, E., editors (2009).
SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09). CEUS-WS.org.

Stolcke, A. (2002).
SRILM - An Extensible Language Modeling toolkit.
In Intl. Conference on Spoken Language Processing, Denver, Colorado.

Taylor, F. (1965).
Cryptomnesia and Plagiarism.
The British Journal of Psychiatry, 111:1111–1118.

Weber, S. (2007).
Das Google-Copy-Paste-Syndrom. Wie Netzplagiate Ausbildung und Wissen gefahrden.
Telepolis.

Whale, G. (1986).
Detection of pagiarism in student programs.
In Proceedings of the Ninth Australasian Computer Science Conference (ACSC 1986), pages 231–241.

Wikipedia (2010a).
Hash.
[Online; accessed 17-Septiembre-2010].

References X

Wikipedia (2010b).
Party of European Socialists | Partido Socialista Europeo | Europako Alderdi Sozialista .
[Online; accessed 10-February-2010].