

Plagiarism Detection

Paolo Rosso

joint work with Alberto Barrón-Cedeño and Enrique Flores
and also with Parth Gupta and Marc Franco-Salvador

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

Natural Language Engineering Lab
PRHLT Research Center
Universitat Politècnica de València, Spain

Text Mining - Diploma on Big Data, ETSINF
20/06/15



LinguaGg
Языковые
Linguaggio
Sprach
Natural Language Engineering Lab
NLEL



Outline

Introduction

Basic Concepts

Intrinsic Plagiarism Detection

External Plagiarism Detection

PAN Task on Plagiarism Detection

Plagiarism and Paraphrasing

Cross-Language Plagiarism Detection

PAN Tasks @ CLEF

Detection of Plagiarism in Source Code

Introduction: Commercial Plagiarism Detection



The Plagiarism Resource Site



Dupli Checker

ephorus



eTBLAST 3.0:
a similarity-based search engine



Moss



Plagiarism Detect.com
THE POWER OF UNIQUENESS



:::plagium™

PlagScan
Your online service for plagiarism detection

safeAssign™
by Blackboard

Viper
The Anti-plagiarism
Scanner

The Sherlock Plagiarism Detector

SID - Plagiarism Detection



YAP

Introduction: A “History” of Plagiarism

Poems kept in the Alexandria's library were presented to a contest by other people. They were judged as thieves



The Roman poet accused Fidentinus, of stealing his verses, calling him *plagiarius* (Latin for kidnapper)



Martial
II A.D.



Gutenberg
XVI A.D.

A literary thief is a plagiarist



Ben Jonson
1601

World's first copyright act (London)



Statute of Anne
1710

A plagiarist is one who steals the thoughts or writings of another"



Samuel Johnson
1755

Imitators only give us a sort of Duplicates of what we had, possibly much better, before. Good authors are original, bad authors copy, and copying is no better than "sordid Theft".



Edward Young
1759



paper mills
computers + Web

XVIII
XX
Alexander Pope



W. Shakespeare



Re-use of history books and other plays
(some of them from Montaigne)

"Copy the Masters instead of inventing"

We have no choice but to steal from the classics because "To copy Nature is to copy them".

[Irribarne and Retondo, 1981, Lynch, 2006]

Introduction: In the news

Daily Mail

JK Rowling sued for £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

Levante
AMERICANAS

A Murcian professor is charged for plagiarising his student thesis

January 29th, 2009

VANGUARDIA

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

October 17th, 2010

Introduction: Plagiarism of Ideas

JK Rowling sued for £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, ”

Introduction: Plagiarism of Ideas

JK Rowling sued for £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

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

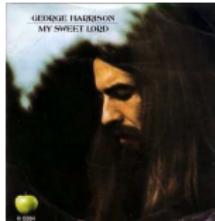
- It can be considered cut-and-paste plagiarism
- It is the easiest to detect

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

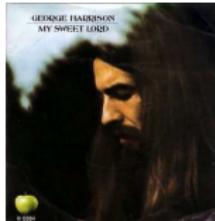


Introduction: Cryptomnesia

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

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: Motivation

- Human beings are the best in detecting a case of re-use, but the explosion in the amount of information causes keeping track of every available resource unaffordable
- The long-term aim is discouraging plagiarism rather than punishing it

$$\text{Approach} = \begin{cases} \text{prevention} \\ \text{surveillance} \\ \text{response} \end{cases}$$

[Maurer et al., 2006]

Introduction: Motivation

- Human beings are the best in detecting a case of re-use, but the explosion in the amount of information causes keeping track of every available resource unaffordable
- The long-term aim is discouraging plagiarism rather than punishing it

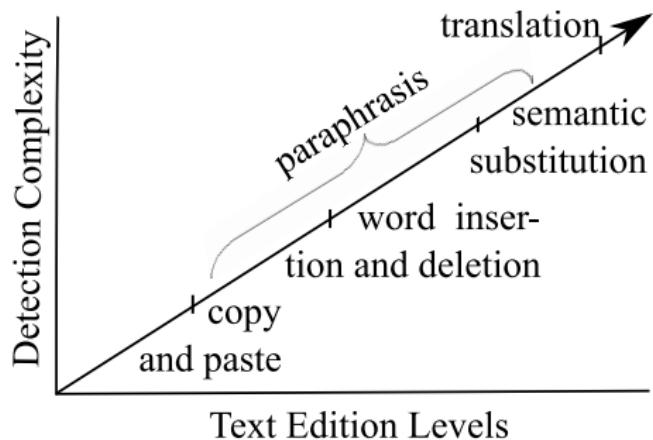
$$\text{Approach} = \begin{cases} \text{prevention} \\ \text{surveillance} \\ \text{response} \end{cases}$$

[Maurer et al., 2006]

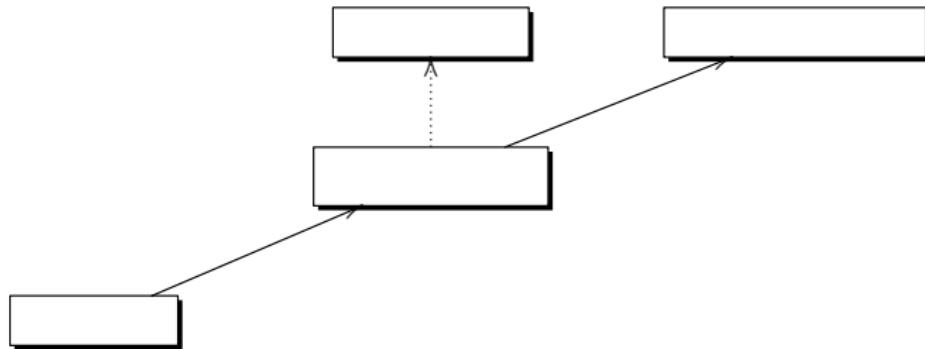
Introduction: Plagiarism Complexity

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

[Maurer et al., 2006]

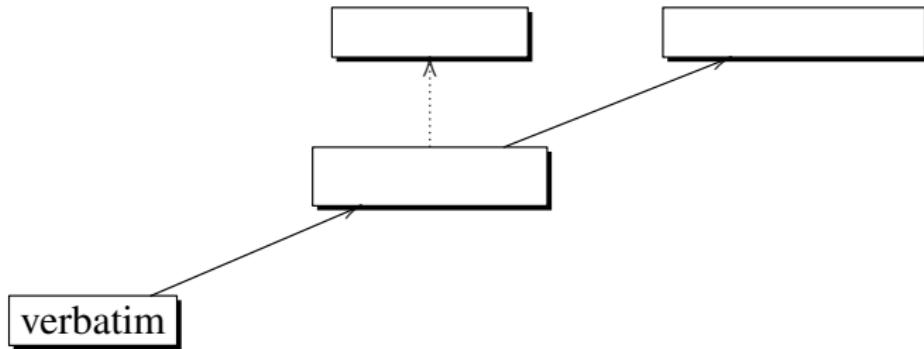


Introduction: Detection Difficulties



[Taylor, 1965, Martin, 1994, Clough et al., 2002, Maurer et al., 2006]

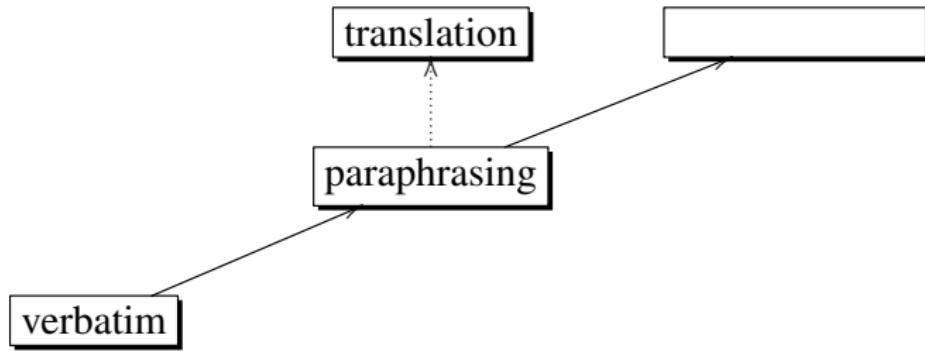
Introduction: Detection Difficulties



relatively easy

[Taylor, 1965, Martin, 1994, Clough et al., 2002, Maurer et al., 2006]

Introduction: Detection Difficulties

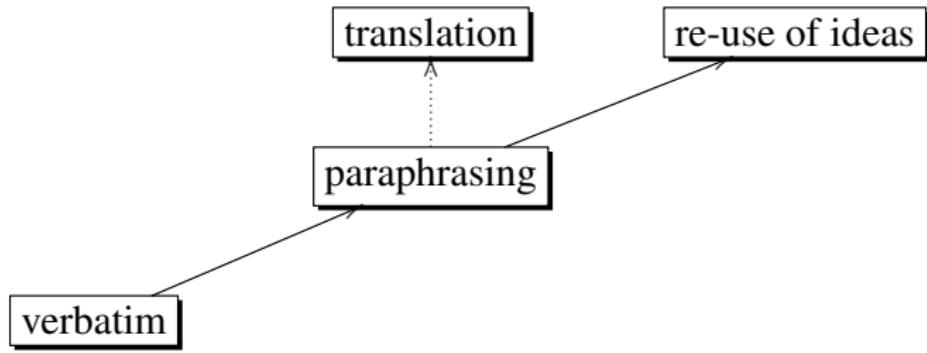


relatively easy

difficult

[Taylor, 1965, Martin, 1994, Clough et al., 2002, Maurer et al., 2006]

Introduction: Detection Difficulties



relatively easy

difficult

nearly impossible

[Taylor, 1965, Martin, 1994, Clough et al., 2002, Maurer et al., 2006]

Introduction: Plagiarism Occurrence

- 33 % (25 %) of students admitted cheating on assignments (quizzes and exams); only 1.3 % reported to have been caught. [Haines et al., 1986]

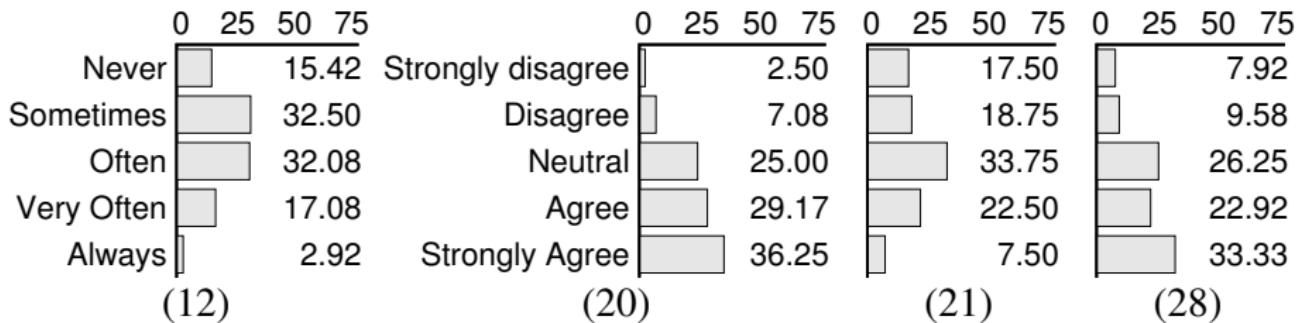
Introduction: Plagiarism Occurrence

- 33 % (25 %) of students admitted cheating on assignments (quizzes and exams); only 1.3 % reported to have been caught. [Haines et al., 1986]
- 80 % of professors declared having found plagiarism in their students works [Chapman and Lupton, 2004]

Introduction: Plagiarism Occurrence

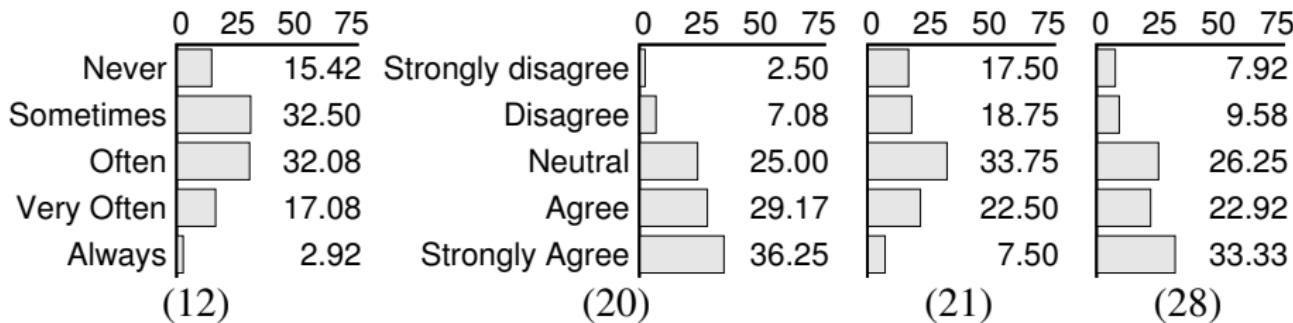
- 33 % (25 %) of students admitted cheating on assignments (quizzes and exams); only 1.3 % reported to have been caught. [Haines et al., 1986]
- 80 % of professors declared having found plagiarism in their students works [Chapman and Lupton, 2004]
- Professors estimate that around 28 % of their pupils' reports include plagiarism [Association of Teachers and Lecturers, 2008]

Introduction: Scholar Practices and Attitudes



12) I include fragments translated from Web pages in my reports

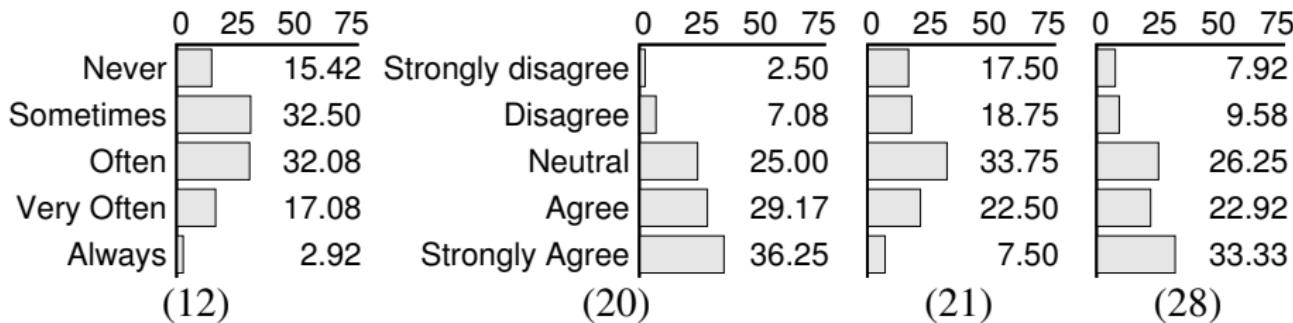
Introduction: Scholar Practices and Attitudes



- ⑫ I include fragments translated from Web pages in my reports
- ⑯ If I express with my own words an idea I have read, I do not need to include any citation and I am not plagiarising

(joint work with Universidad Tecnológica del Valle de Toluca) □ [Barrón-Cedeño, 2012] ↗ ↘ ↙

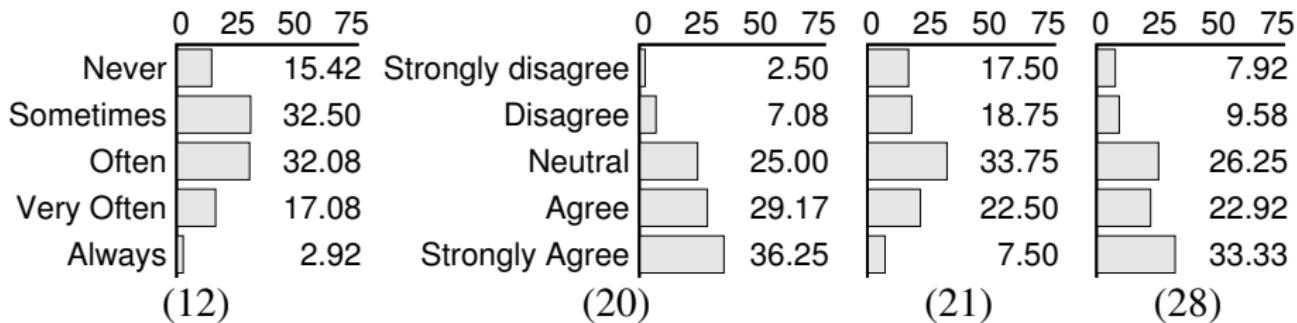
Introduction: Scholar Practices and Attitudes



- ⑫ I include fragments translated from Web pages in my reports
- ⑯ If I express with my own words an idea I have read, I do not need to include any citation and I am not plagiarising
- ㉑ Translating a text fragment and including it into my report is not plagiarism, as the words are different

(joint work with Universidad Tecnológica del Valle de Toluca) □ [Barrón-Cedeño, 2012] ↗ ↘ ↙

Introduction: Scholar Practices and Attitudes



- ⑫ I include fragments translated from Web pages in my reports
- ⑯ If I express with my own words an idea I have read, I do not need to include any citation and I am not plagiarising
- ㉑ Translating a text fragment and including it into my report is not plagiarism, as the words are different
- ㉘ Among all of the websites, Wikipedia is one of the most frequently used for plagiarising

(joint work with Universidad Tecnológica del Valle de Toluca) □ [Barrón-Cedeño, 2012] ↗ ↘ ↙ ↘

Introduction: Is it 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 it 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)

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: Text Re-Use and Plagiarism

- Text re-use is “the situation in which pre-existing written material is consciously used again during the creation of a new text or version”
[Clough and Gaizauskas, 2009]

Instances: summarisation, translation, simplification

Introduction: Text Re-Use and Plagiarism

- Text re-use is “the situation in which pre-existing written material is consciously used again during the creation of a new text or version”
[Clough and Gaizauskas, 2009]

Instances: summarisation, translation, simplification

- It can become plagiarism...

by reference omission

of authorship

Introduction: Text Re-Use and Plagiarism

- Text re-use is “the situation in which pre-existing written material is consciously used again during the creation of a new text or version”
[Clough and Gaizauskas, 2009]

Instances: summarisation, translation, simplification

- It can become plagiarism...

by reference omission

of authorship

by inappropriate quotation

by amount of borrowing

[Martin, 1994, Maurer et al., 2006]

Introduction: Text Re-Use and Plagiarism

- Text re-use is “the situation in which pre-existing written material is consciously used again during the creation of a new text or version”
[Clough and Gaizauskas, 2009]

Instances: summarisation, translation, simplification

- It can become plagiarism...

by reference omission

of authorship

by inappropriate quotation

by amount of borrowing

of secondary sources

[Martin, 1994, Maurer et al., 2006]

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]

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]

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
newspapers
Wikis (Wikipedia)
collaborative authoring

Plagiarism
students reports
web contents
scientific papers

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
newspapers
Wikis (Wikipedia)
collaborative authoring

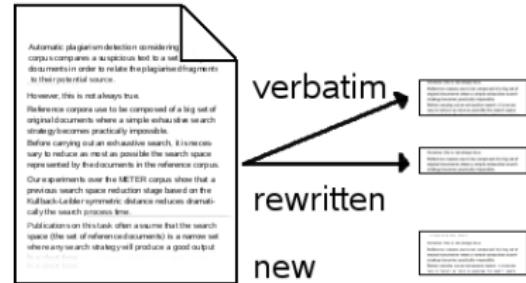
Plagiarism
students reports
web contents
scientific papers

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

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 Telegraph version

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
<p>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.</p>	<p>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.</p>

Introduction: The METER Project

PA version	The Telegraph version
<p>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.</p>	<p>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.</p>

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

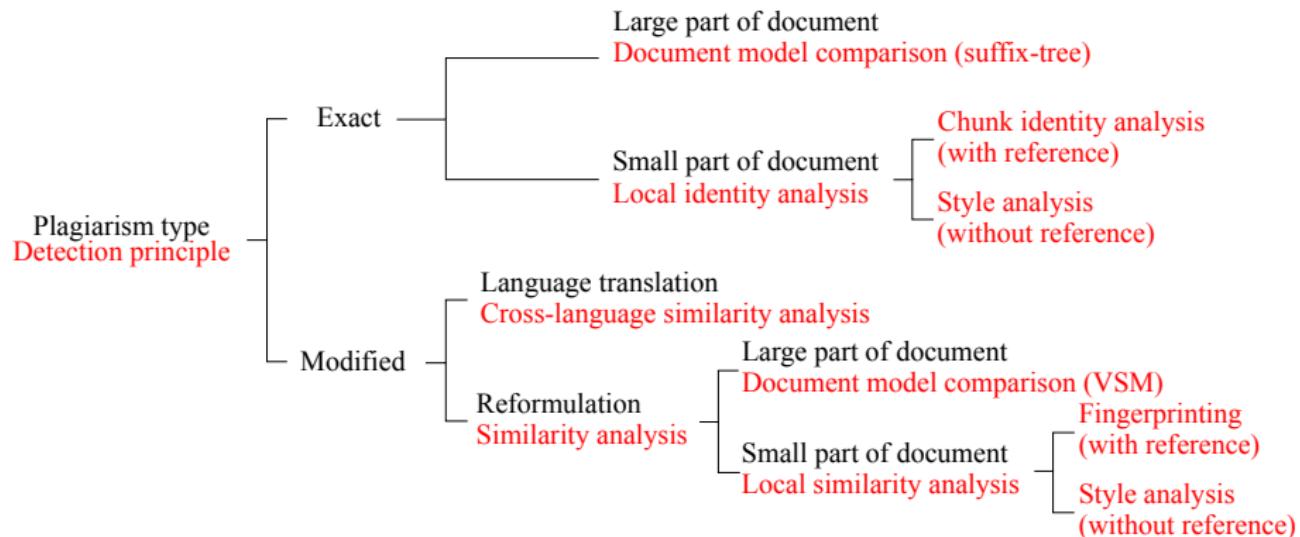
Given a (set of) suspicious document(s) and a set of source documents, find all plagiarised sections in the suspicious document(s) and, if available, the corresponding source sections.

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

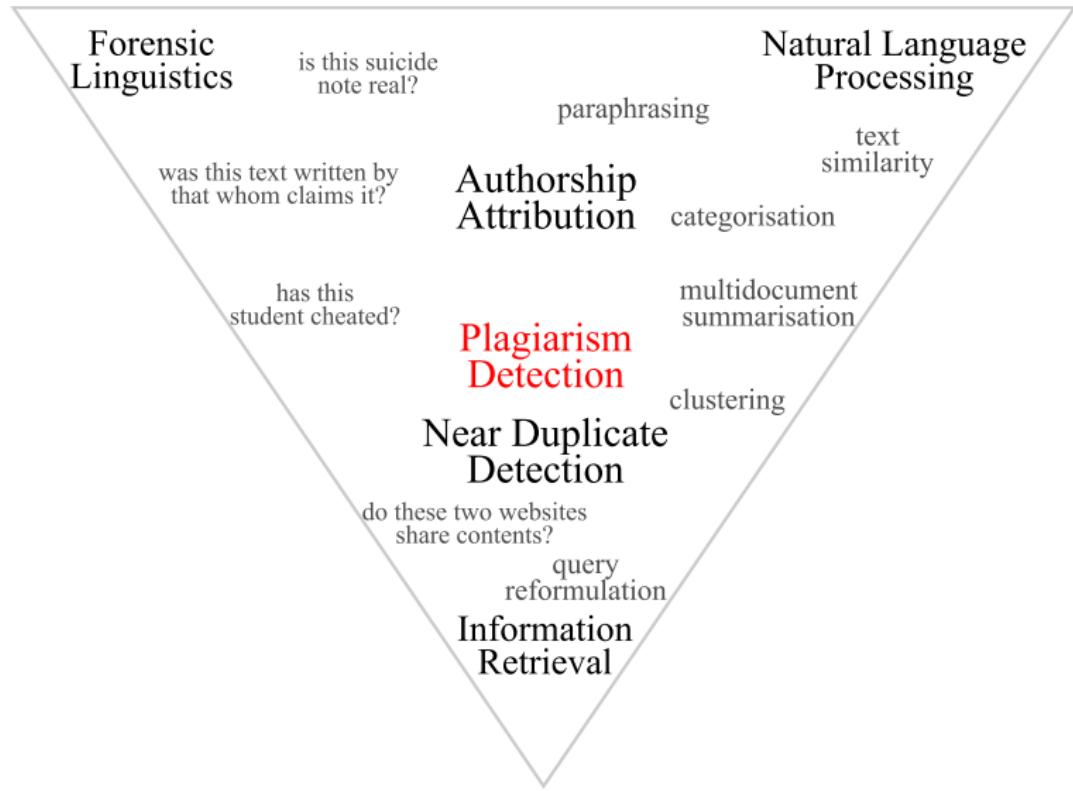


[Meyer zu Eißen and Stein, 2006]

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

Introduction

Basic Concepts

Intrinsic Plagiarism Detection

External Plagiarism Detection

PAN Task on Plagiarism Detection

Plagiarism and Paraphrasing

Cross-Language Plagiarism Detection

PAN Tasks @ CLEF

Detection of Plagiarism in Source Code

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

The diagram shows the word "example" with its characters underlined in pairs. To the right, four 3-grams are listed in a grid:

exa	xam
amp	mpl
ple	

- word 2-grams

this is just
an example

The diagram shows the sentence "this is just an example" with its words underlined in pairs. To the right, four 2-grams are listed in a grid:

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]

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:

- $md5sum(\text{this is a test}) = e19c1283c925b3206685ff522acf3e6$
- $RabinKarp(\text{starwarsisanepicsspaceoperafranchiseinitiallyconcei}) = 4742204955$

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)

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)

$$I_G(\text{comic}) = 6$$

$$I_G(\text{Newsweek}) = 10$$

$$I_G(T_1) = 15.2$$

$$I_G(T_2) = 14.1$$

(also Flesch–Kincaid readability test, among others)

Basics: Word Frequency Class

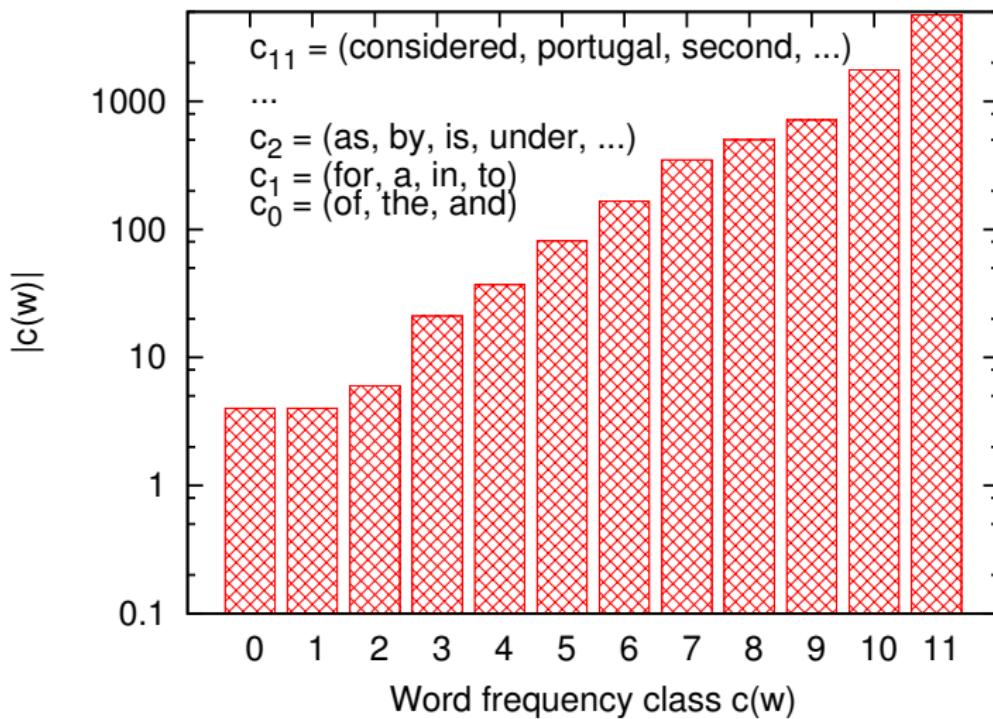
Given the corpus \mathcal{D} , the word frequency class is defined as:

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

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

Basics: Similarity Measures

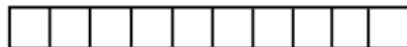
$$\text{sim}(d, d_q) \in [0, 1]$$

- $\text{sim}(d, d_q) = 0 \rightarrow d$ and d_q are not similar at all
- $\text{sim}(d, d_q) = 1 \rightarrow d$ 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_q 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					d_q			
,	epic	it	star		,	distinguish	in	sequels	
.	expanded	its	subtitle		:	epic	initially	simply	
1970s	film	later	that		.	episodeiv	is	space	
a	first	lucas	the		a	film	it	star	
added	franchise	new	time		added	first	its	subtitle	
an	from	opera	titled		an	franchise	later	the	
and	george	prequels	to		and	from	lucas	titled	
but	had	sequels	wars		but	george	new	to	
by	hope	significantly	was		by	had	opera	wars	
conceived	in	simply			conceived	hope	prequels	was	
distinguish	initially	since							
during	is	space							

Basics: Similarity Measures - VSM

Jaccard Coefficient

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

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

Basics: Similarity Measures - VSM

Cosine Similarity

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

ω_t is estimated by the well known *tf*

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

Basics: Similarity Measures - VSM

Cosine Similarity

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

Basics: Similarity Measures - VSM

Cosine Similarity

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

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

Basics: Similarity Measures - VSM

Word Chunking Overlap

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

- Based on the so called asymmetric subset measure:

$$subset(d, d_q) = \frac{\sum_{t_i \in c(d, d_q)} tf_{t,d} \cdot tf_{t,d_q}}{\sum_{t_i \in d} tf_{t_i,d}^2}$$

Basics: Similarity Measures - VSM

Word Chunking Overlap

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

- Based on the so called asymmetric subset measure:

$$\text{subset}(d, d_q) = \frac{\sum_{t_i \in c(d, d_q)} t f_{t,d} \cdot t f_{t,d_q}}{\sum_{t_i \in d} t f_{t_i,d}^2}$$

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

$$\varepsilon - \left(\frac{t f_{t,d}}{t f_{t,d_q}} + \frac{t f_{t,d_q}}{t f_{t,d}} \right) > 0$$

[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, $\varepsilon = 2.5$)

[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, $\varepsilon = 2.5$)

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

[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, $\varepsilon = 2.5$)

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

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

Basics: Similarity Measures - VSM

Word Chunking Overlap

- By considering $\varepsilon = 2.5$

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

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

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

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

[Schleimer et al., 2003]

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

[Schleimer et al., 2003]

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

starwarsisanepicspaceoperafra
nchiseinitiallyconceivedbygeorg
elucasduringthe1970

d_q

starwarsisanepicspaceope
rafranchiseinitiallyconceive
dbygeorgelucas

Basics: Similarity Measures - Fingerprinting

Winnowing

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

Basics: Similarity Measures - Fingerprinting

Winnowing

d
starwarsisanepicspaceoperafra nchiseinitiallyconceivedbygeorg elucasduringthe1970

[4742204955 4690954177 51549901
624610790 -2470793273 -1315199375
3953400264 -78415511 664863318
3374288481 4230663014 -3213422081
-2056259009 7513105677 -6553730326
5257922027 4828416784 -8476824670
9011767372 1240867252]

d_q
starwarsisanepicspaceope rafranchiseinitiallyconceive dbygeorgelucas

[4742204955 4690954177
51549901 624610790
-2470793273 -1315199375
3953400264]

By considering $t = 20$

$$sim(d, d_q) = \frac{\emptyset}{2} = 0$$

Basics: Similarity Measures - Fingerprinting

Winnowing

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

d_q
starwarsisanepicspaceope rafranchiseinitiallyconceive dbygeorgelucas

By considering $t = 20$

By considering $t = 10$

$$\text{sim}(d, d_q) = \frac{\emptyset}{2} = 0$$

$$\text{sim}(d, d_q) = \frac{1}{3}$$

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.

[Bernstein and Zobel, 2004]

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 ($\max(n) = 8$).

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

Basics: Similarity Measures - Fingerprinting

SPEX

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

Basics: Similarity Measures - Fingerprinting

SPEX

$n = 2$	d		d_q
star wars	in the	star wars	in the
wars is	the franchise	wars is	the franchise
is an	franchise was	is an	franchise was
an epic	was simply	an epic	was simply
epic space	simply titled	epic space	simply titled
space opera	titled star	space opera	titled star
opera franchise	star wars	opera franchise	star wars
franchise initially	wars but	franchise initially	wars but
initially conceived	but later	initially conceived	but later
conceived by	later had	conceived by	later had
by george	had the	by george	had the
george lucas	the subtitle	george lucas	the subtitle
the first	subtitle a	lucas the	subtitle a
first film	a new	the first	a new
film in	new hope	first film	new hope
		film in	

Basics: Similarity Measures - Fingerprinting

SPEX

$n = 2$	d		d_q
star wars	in the	star wars	in the
wars is	the franchise	wars is	the franchise
is an	franchise was	is an	franchise was
an epic	was simply	an epic	was simply
epic space	simply titled	epic space	simply titled
space opera	titled star	space opera	titled star
opera franchise	star wars	opera franchise	star wars
franchise initially	wars but	franchise initially	wars but
initially conceived	but later	initially conceived	but later
conceived by	later had	conceived by	later had
by george	had the	by george	had the
george lucas	the subtitle	george lucas	the subtitle
the first	subtitle a	lucas the	subtitle a
first film	a new	the first	a new
film in	new hope	first film	new hope
		film in	

Basics: Similarity Measures - Fingerprinting

SPEX

$n = 3$	d	d_q
star wars is	in the franchise	star wars is
wars is an	the franchise was	wars is an
is an epic	franchise was s...	is an epic
an epic space	was simply titled	an epic space
epic space opera	simply titled star	epic space opera
space opera fran...	titled star wars	space opera fra...
opera franchise in...	star wars but	opera franchise in...
franchise initially ...	wars but later	franchise initially ...
initially conceived ...	but later had	initially conceived ...
conceived by g...	later had the	conceived by g...
by george lucas	had the subtitle	by george lucas
the first film	the subtitle a	the first film
first film in	subtitle a new	first film in
film in the	a new hope	film in the

Basics: Similarity Measures - Fingerprinting

SPEX

- By considering $n = 3$ (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 $[0, 1]$

Machine Translation

- Given a text e written in a language L , to find the most likely translation f , in a language L'

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

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

Basics: Similarity Measures - Probabilistic

Machine Translation. IBM Model Adaptation

$$\text{sim}(d, d_q) = \varrho(d) w(d_q | d)$$

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

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

Basics: Similarity Measures - Probabilistic

Machine Translation. IBM Model Adaptation

$$\text{sim}(d, d_q) = \varrho(d) w(d_q | d)$$

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

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

- $p(x, y)$ 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

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

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

Basics: Similarity Measures - Probabilistic

Machine Translation

$n = 1$	d	d_q
star	during	star
wars	the	wars
is	1970s	is
an	and	an
epic	significantly	epic
space	expanded	space
opera	since	opera
franchise	that	franchise
initially	time	initially
conceived	the	conceived
by	first	by
george	film	george
lucas	in	lucas
	new	
	hope	

Basics: Similarity Measures - Probabilistic

Machine Translation

$n = 1$	d	d_q
star	during	star
wars	the	wars
is	1970s	is
an	and	an
epic	significantly	epic
space	expanded	space
opera	since	opera
franchise	that	franchise
initially	time	initially
conceived	the	conceived
by	first	by
george	film	george
lucas	in	lucas
	new	
	hope	

Basics: Similarity Measures - Probabilistic

Machine Translation

$n = 1$	d	d_q
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

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

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

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

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

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

- 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
- $KL_\delta(P_{d_q} \parallel Q_d) = 0 \rightarrow P_{d_q} = Q_d$ and the documents are quite similar.

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

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

Basics: Similarity Measures - Probabilistic

Kullback-Leibler

keywords in d ranked by $tf-idf$			P_{d_q}
1970s	expanded	new	1970s 0.01886
lucas	titled	but	lucas 0.01886
george	conceived	was	george 0.01886
star	initially	had	star 0.03773
wars	film	is	wars 0.03773
epic	simply	that	epic 0.01886
franchise	during	an	franchise 0.03773
subtitle	since	and	
hope	later	by	
opera	time	a	
subtitle	space	in	
hope	first	the	

Basics: Similarity Measures - Probabilistic

Kullback-Leibler

keywords in d ranked by $tf-idf$			P_{d_q}
1970s	expanded	new	1970s 0.01886
lucas	titled	but	lucas 0.01886
george	conceived	was	george 0.01886
star	initially	had	star 0.03773
wars	film	is	wars 0.03773
epic	simply	that	epic 0.01886
franchise	during	an	franchise 0.03773
subtitle	since	and	Q_d
hope	later	by	1970s 0.0002
opera	time	a	lucas 0.0216
subtitle	space	in	george 0.0216
hope	first	the	star 0.0433
			wars 0.0433
			epic 0.0213
			franchise 0.0433

Basics: Similarity Measures - Probabilistic

Kullback-Leibler distance

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

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

Basics: Similarity Measures - Probabilistic

Okapi BM25

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

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

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

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

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

- $k_1 = 0$ 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 $b = 0.75$

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

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

- $k_1 = 0$ 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 .

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 .

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

Basics: Plagiarism Detection

- a text is read and certain characteristics found which suggest plagiarism
- possible source texts of a suspicion fragment are searched using tools such as Web search engines or non-digital material

(adapted from [Clough, 2003, Meyer zu Eißen and Stein, 2006,
Barrón-Cedeño et al., 2008, Potthast et al., 2009])

Basics: Plagiarism Detection

- a text is read and certain characteristics found which suggest plagiarism
- possible source texts of a suspicion fragment are searched using tools such as Web search engines or non-digital material

Intrinsic plagiarism detection Let d_q be presumably written by \mathcal{A} .

Determine whether the contents in d_q were actually written by \mathcal{A} . If not, extract those fragments potentially written by a \mathcal{A}'

(adapted from [Clough, 2003, Meyer zu Eißen and Stein, 2006,
Barrón-Cedeño et al., 2008, Potthast et al., 2009])

Basics: Plagiarism Detection

- a text is read and certain characteristics found which suggest plagiarism
- possible source texts of a suspicion fragment are searched using tools such as Web search engines or non-digital material

Intrinsic plagiarism detection Let d_q be presumably written by \mathcal{A} .

Determine whether the contents in d_q were actually written by \mathcal{A} . If not, extract those fragments potentially written by a \mathcal{A}'

external plagiarism detection Let d_q be a suspicious text.

Let D be a set of potential source texts. Determine whether d_q contains borrowed text from a specific $d \in D$.

(adapted from [Clough, 2003, Meyer zu Eißen and Stein, 2006,
Barrón-Cedeño et al., 2008, Potthast et al., 2009])

Basics: Plagiarism Detection

- a text is read and certain characteristics found which suggest plagiarism
- possible source texts of a suspicion fragment are searched using tools such as Web search engines or non-digital material

Intrinsic plagiarism detection Let d_q be presumably written by \mathcal{A} .

Determine whether the contents in d_q were actually written by \mathcal{A} . If not, extract those fragments potentially written by a \mathcal{A}'

Cross-language external plagiarism detection Let d_q be a suspicious text.

Let D be a set of potential source texts. Determine whether d_q contains borrowed text from a specific $d \in D$. $d_q \in L$, $d' \in L'$ ($L \neq L'$); potential borrowings after translation

(adapted from [Clough, 2003, Meyer zu Eißen and Stein, 2006,
Barrón-Cedeño et al., 2008, Potthast et al., 2009])

Outline

Introduction

Basic Concepts

Intrinsic Plagiarism Detection

External Plagiarism Detection

PAN Task on Plagiarism Detection

Plagiarism and Paraphrasing

Cross-Language Plagiarism Detection

PAN Tasks @ CLEF

Detection of Plagiarism in Source Code

Intrinsic Plagiarism Detection

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

Plagiarism is considered as one of another author's work is considered as one of the plagiarist's own work. If a plagiarist copies a sentence from another author's work and plagiarizes it 10 times, ten plagiarism is observed at an inappropriate time, with the same or different words.

Most of the time, many may be Web pages full of such copied content, images, sounds, and videos easily accessible, that is, can be plagiarized.

Plagiarism detection is the automatic identification of plagiarism and the detection of the original source. It is a hot and developing research field in computer science and information systems. It has been applied in many areas of application in the field of scientific research, business, teaching, and education, among others.

The manual analysis of text with respect to plagiarism becomes infeasible as a large scale, as that automatic plagiarism detection becomes unavoidable.

We present in this paper a system for the detection of plagiarism, namely Cross-language plagiarism detection. This paper is organized as follows. In Section 2, we introduce the basic concepts of writing. In Section 3, we introduce the concept of cross-language plagiarism detection. In Section 4, we introduce the proposed system. In Section 5, we present the experimental results. Finally, in Section 6, we draw our conclusions. In the last section, we discuss the future work.

There are no studies which discuss the amount of cross-language plagiarism directly, but in [21] the authors argue that the number of detected plagiarisms is proportional to the number of plagiarized documents, which was also the case in our experiments.

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

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

Intrinsic Plagiarism Detection

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

Plagiarism is considered as one author's work is considered as one of the plagiaristic copy. If the original text is copied without any changes, it is called direct plagiarism. If the original text is copied and modified or altered, it is called indirect plagiarism. If the original text is copied and used in a different context, it is called derivative plagiarism.

Plagiarism detection is the process of detecting plagiarized text from a large collection of documents.

Plagiarism detection is the automatic identification of plagiarism and the retrieval of the original source. It is a hard and difficult task.

There are many types of plagiarism, such as direct, indirect, derivative, and self-plagiarism.

The main goal of plagiarism detection is to identify plagiarized text and to find the original source.

Plagiarism detection is a challenging task because it requires a large amount of data and complex algorithms.

There are many methods for plagiarism detection, such as statistical, machine learning, and deep learning methods.

Plagiarism detection is an important task in the field of text mining, especially in the field of academic research.

Plagiarism detection is a challenging task because it requires a large amount of data and complex algorithms.

Plagiarism detection is an important task in the field of text mining, especially in the field of academic research.

Plagiarism detection is an important task in the field of text mining, especially in the field of academic research.

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

In this work, we have carried out some research on the influence that mineral salts on the mood of people. For this research I have worked with 5 people who have taken water with different amount of mineral salts. Our theory is that the more minerals are in the water, the more moody people are. [...]

Mineral salts are inorganic molecules of easy ionization in presence of water; in living beings they appear by precipitation as well as dissolved mineral salts. [...] Dissolved mineral salts are always ionized. These salts have a structural function and pH regulating functions, of the osmotic pressure and of biochemical reactions, in which specific ions are involved.

It seems to me that the results are good. [...]

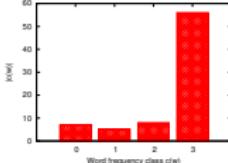
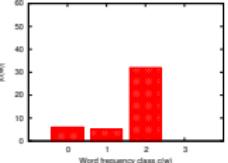
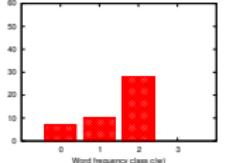
Intrinsic Plagiarism Detection

In this work, we have carried out some research on the influence that mineral salts on the mood of people. For this research I have worked with 5 people who have taken water with different amount of mineral salts. Our theory is that the more minerals are in the water, the more moody people are. [...]

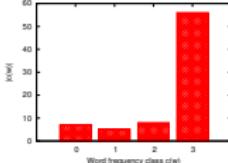
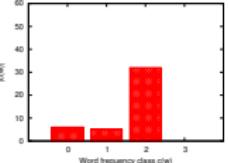
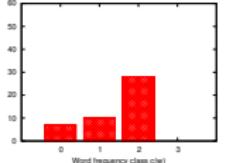
Mineral salts are inorganic molecules of easy ionization in presence of water; in living beings they appear by precipitation as well as dissolved mineral salts. [...] Dissolved mineral salts are always ionized. These salts have a structural function and pH regulating functions, of the osmotic pressure and of biochemical reactions, in which specific ions are involved.

It seems to me that the results are good. [...]

Intrinsic Plagiarism Detection

Measure	Global	■	■
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. measure	16.72	17.07	13.82

Intrinsic Plagiarism Detection

Measure	Global	■	■
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. measure	16.72	17.07	13.82



Stylysis: <http://memex2.dsic.upv.es/StylisticAnalysis>

Outline

Introduction

Basic Concepts

Intrinsic Plagiarism Detection

External Plagiarism Detection

PAN Task on Plagiarism Detection

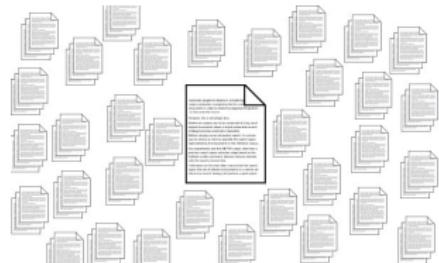
Plagiarism and Paraphrasing

Cross-Language Plagiarism Detection

PAN Tasks @ CLEF

Detection of Plagiarism in Source Code

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

[Potthast et al., 2009]

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 make this task difficult

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

[Potthast et al., 2009]

External Plagiarism Detection

Issues that make 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]

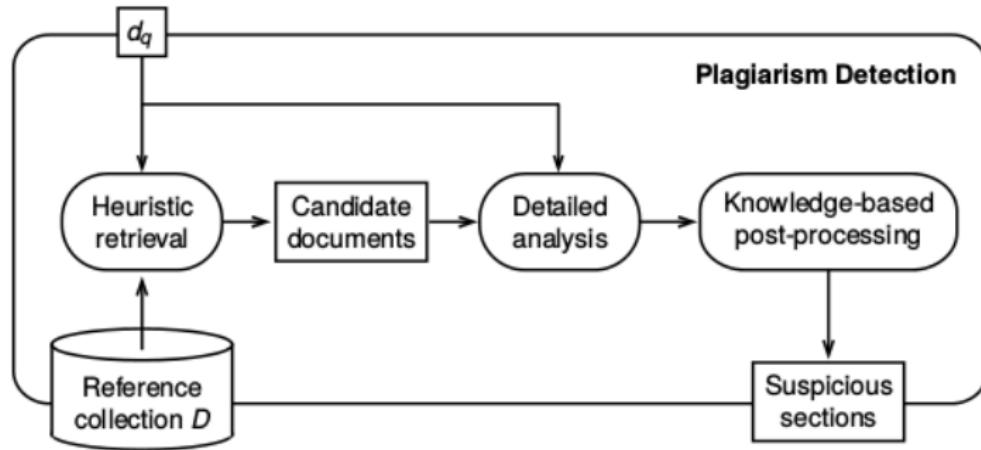
Fingerprinting techniques

SPEX [Bernstein and Zobel, 2004]

Winnowing [Schleimer et al., 2003]

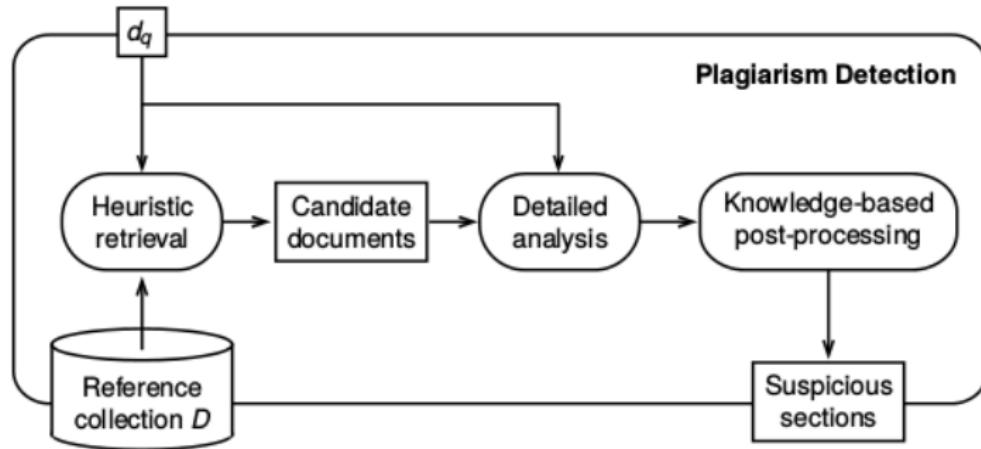
[Potthast et al., 2009]

External: External Prototypical Steps



[Stein et al., 2007]

External: External Prototypical Steps



[Stein et al., 2007]

(we proposed a model for heuristic retrieval based on the Kullback-Leibler distance and the selection of 10 % of a document's vocabulary
[Barrón-Cedeño et al., 2009])

External: Countermeasures

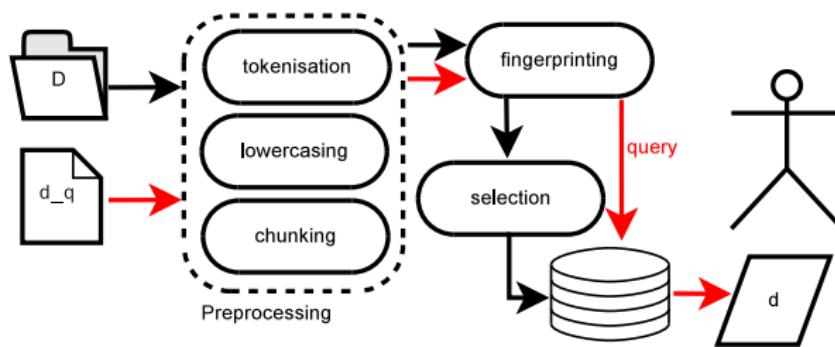
source Copying words or ideas from someone else without giving credit.

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

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

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”

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.

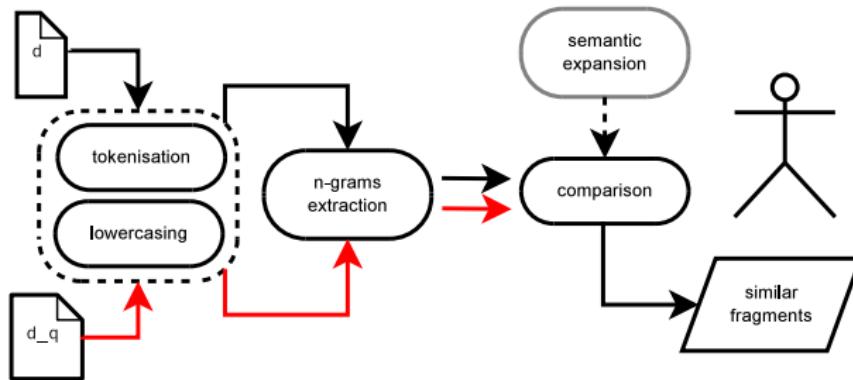
modified copy **Copying the words** and **ideas from someone else's text without giving credit.**

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

External: 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 $s_{\{1 \dots i \dots I\}}$
- $N(s_i)$ is the set of n -grams in s_i

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 $s_{\{1 \dots i \dots I\}}$
- $N(s_i)$ is the set of n -grams in s_i
- The containment measure (cosine or Jaccard coefficient) can be calculated [Broder, 1997]

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

[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

[Barrón Cedeño, 2008]

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

[Barrón Cedeño, 2008]

External: n -grams

Why n -grams work?

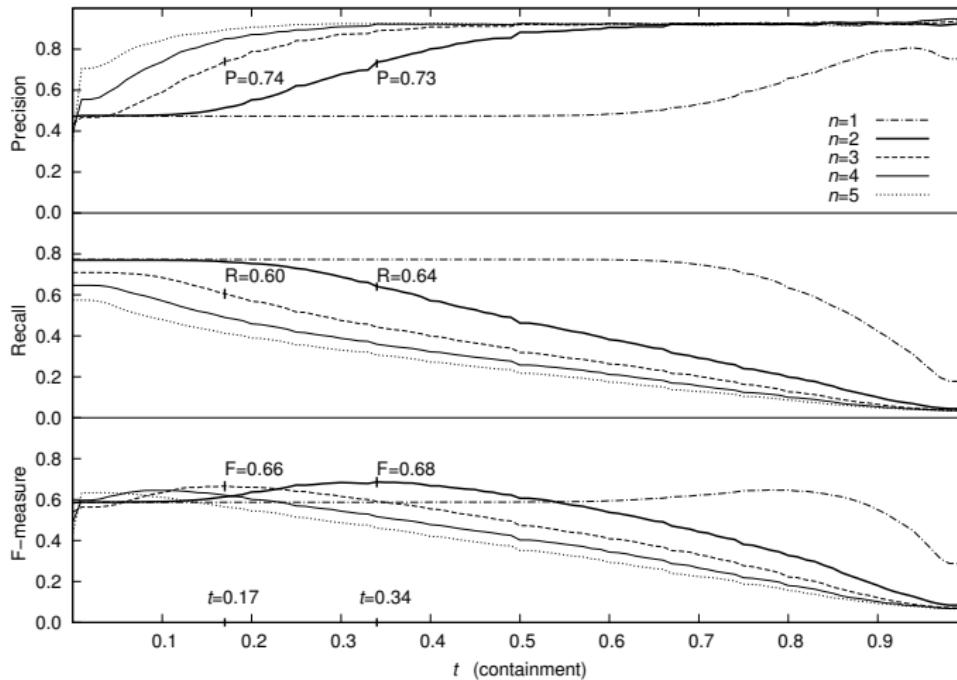
- 4 documents (3,728 words in average)
- One author \mathcal{A}
- One topic

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

Exercise: Increase n until getting a hapax legomena on the Web

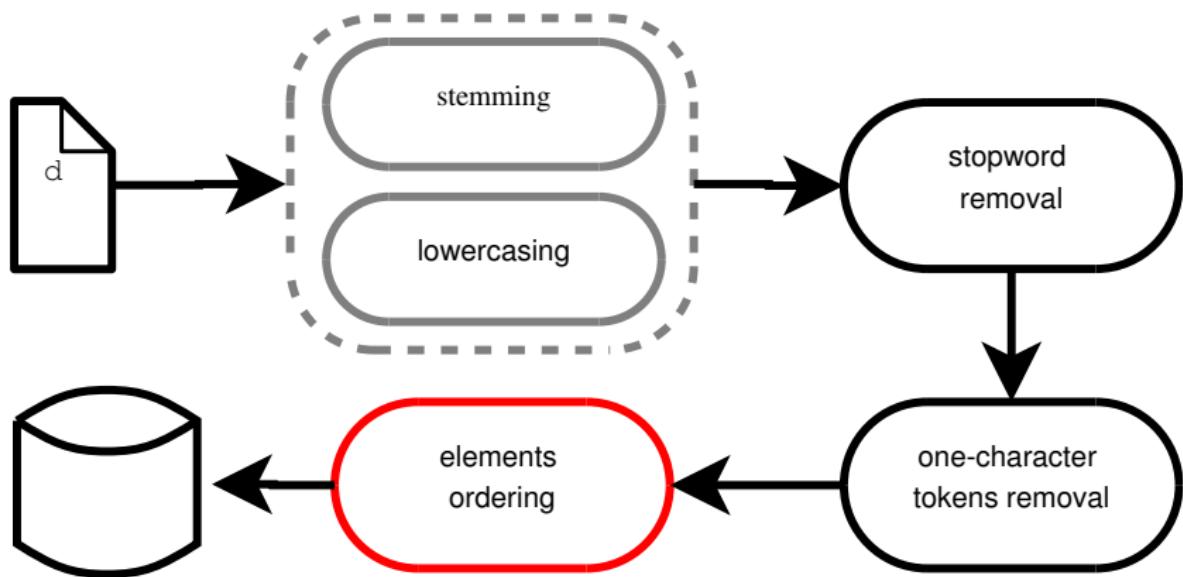
[Barrón Cedeño, 2008]

External: Definition of n (METER Corpus)



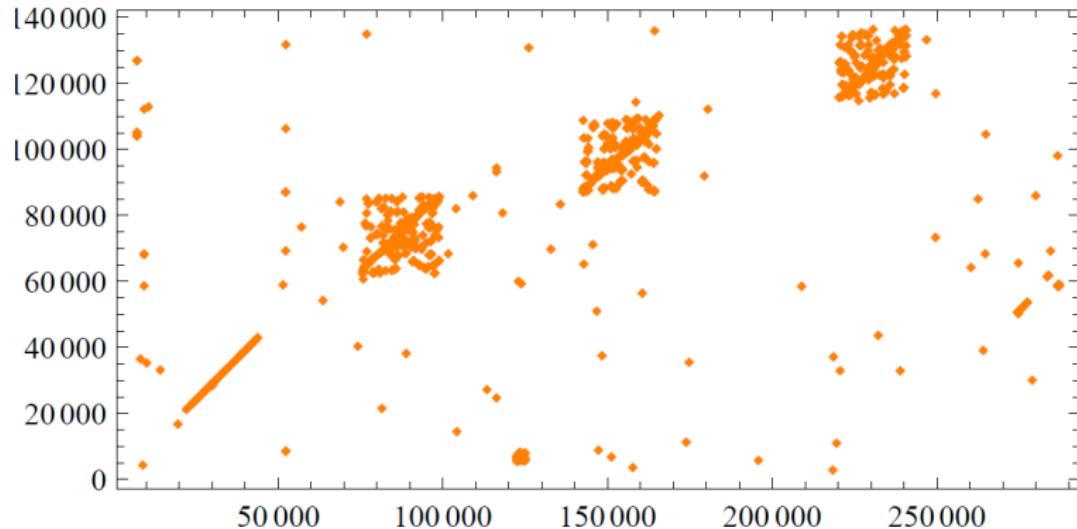
[Barrón-Cedeño and Rosso, 2009]

External: Contextual n -grams



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

External: Dotplot techniques



Documents n -grams are in x and y . A dot means that the n -gram exists in both documents.

[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)

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 | hypernym |

The **peculiar** incident of the **pet** in the night

[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: $c_a, c_b, c_c, \dots, c_z$
- A standard distribution of index term frequencies can be stated (for instance, from the British National Corpus)
- 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 Fuzzification of Δ_{pf}

External: Fuzzy Fingerprinting

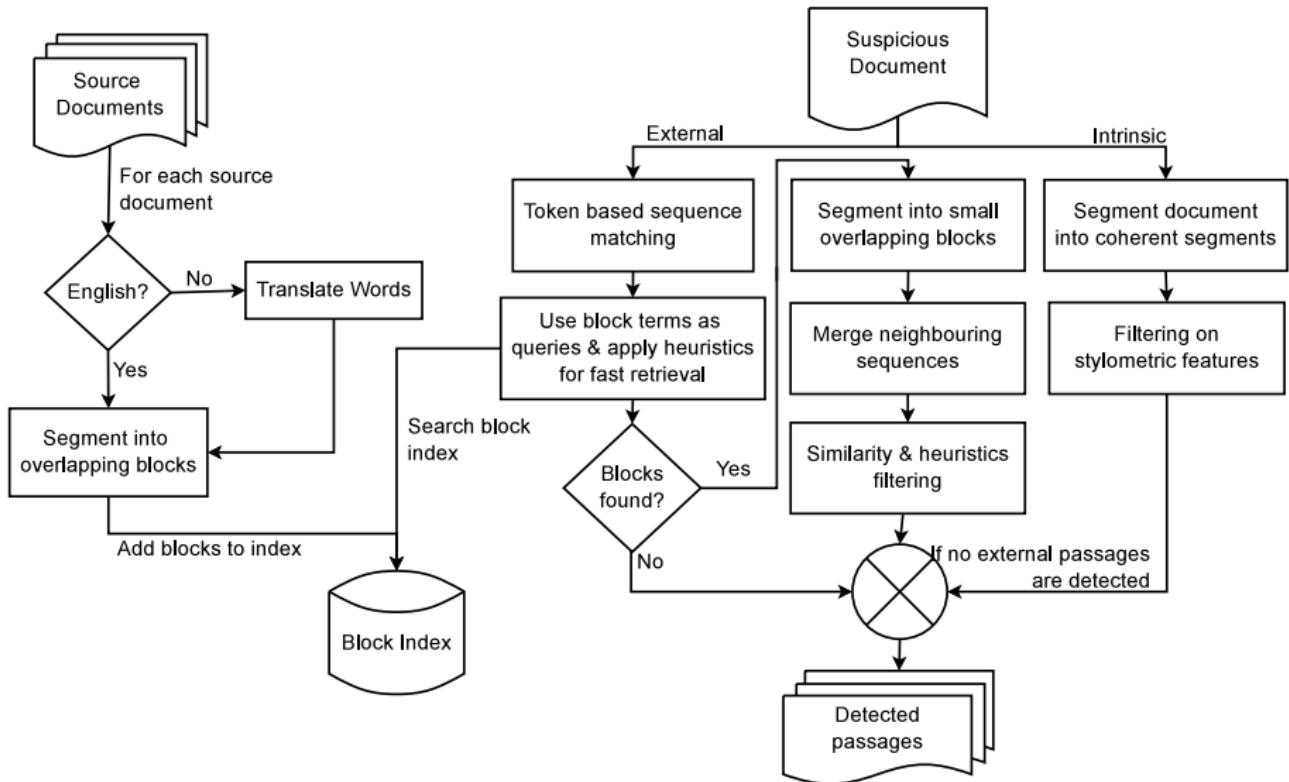
The fuzzy fingerprint $h_\varphi(d)$ is constructed within the following steps:

- 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 Fuzzification of Δ_{pf}

Hash collision

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

External: IR Approach



[Muhr et al., 2010]

Outline

Introduction

Basic Concepts

Intrinsic Plagiarism Detection

External Plagiarism Detection

PAN Task on Plagiarism Detection

Plagiarism and Paraphrasing

Cross-Language Plagiarism Detection

PAN Tasks @ CLEF

Detection of Plagiarism in Source Code



<http://pan.webis.de>

Lab on Uncovering Plagiarism, Authorship, and Social Software Misuse

PAN-PC-09: Corpus of Synthetic Plagiarism

- Plagiarism implies an ethical issue
- Nobody would like to be included in a corpus containing plagiarism!

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

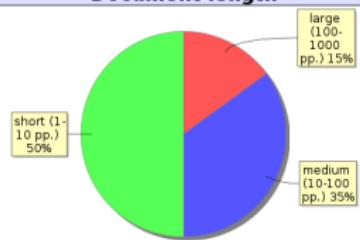
<http://www.webis.de/research/corpora>

PAN-PC-09: Corpus Parameters

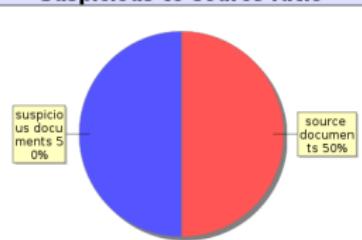
- Document length
- Suspicious-to-source ratio
- Plagiarism percentage
- Cases length
- Plagiarism language
- Cases obfuscation

PAN-PC-09: Corpus Parameters

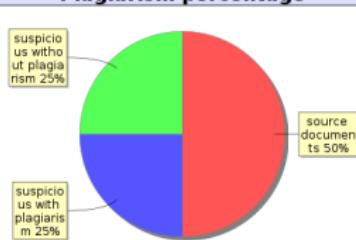
Document length



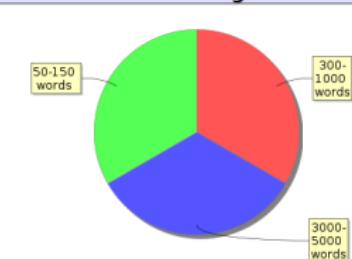
Suspicious-to-source ratio



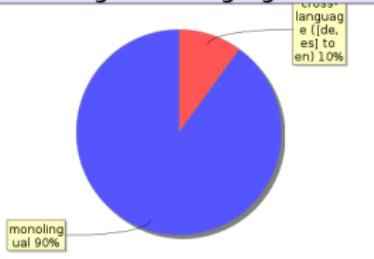
Plagiarism percentage



Cases length



Plagiarism Languages



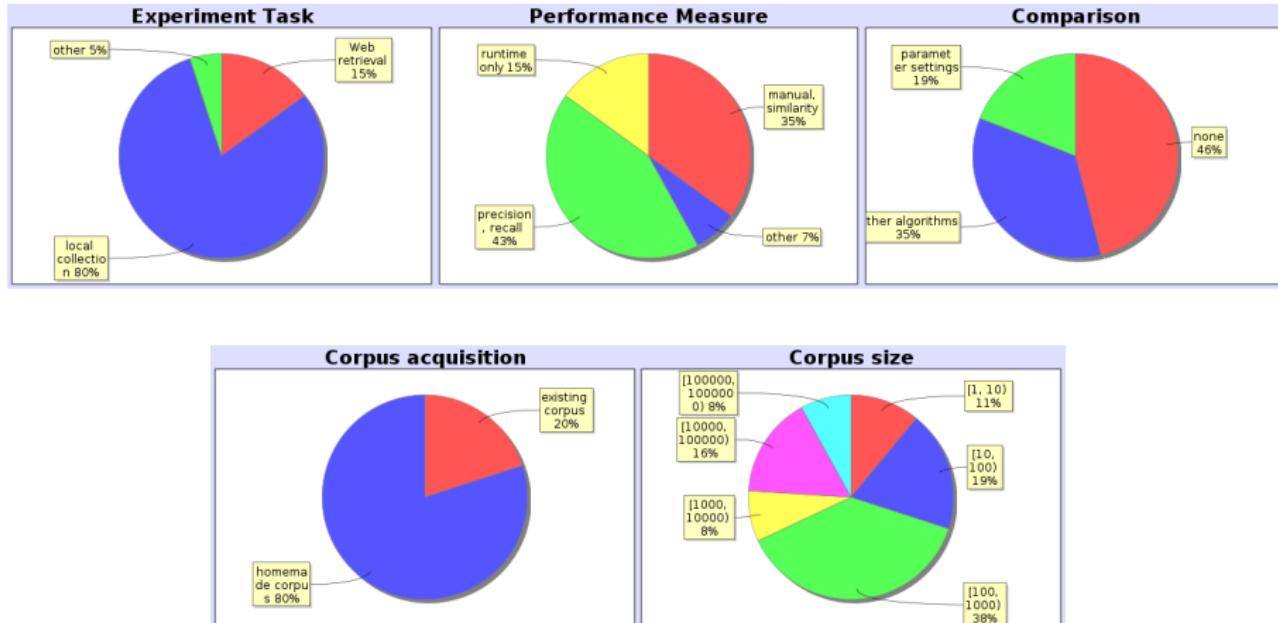
PAN-PC-09: Simulating Obfuscation

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., 2010]

PAN: How Researchers Evaluate Plag. Detection

- No standard evaluation measures have been previously defined
- Evaluations used to be incomparable and often not even reproducible

PAN: How Researchers Evaluate Plag. Detection

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

PAN: Evaluation Measures

We are interested in considering three main aspects when evaluating a Plagiarism Detection Algorithm (PDA):

- ① plagiarised and—if available—source fragments are retrieved;
- ② original text fragments are not reported as plagiarised; and
- ③ plagiarised fragments are not detected over and over again.

PAN: Evaluation Measures

Precision and Recall

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

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

PAN: Evaluation Measures

Precision and Recall

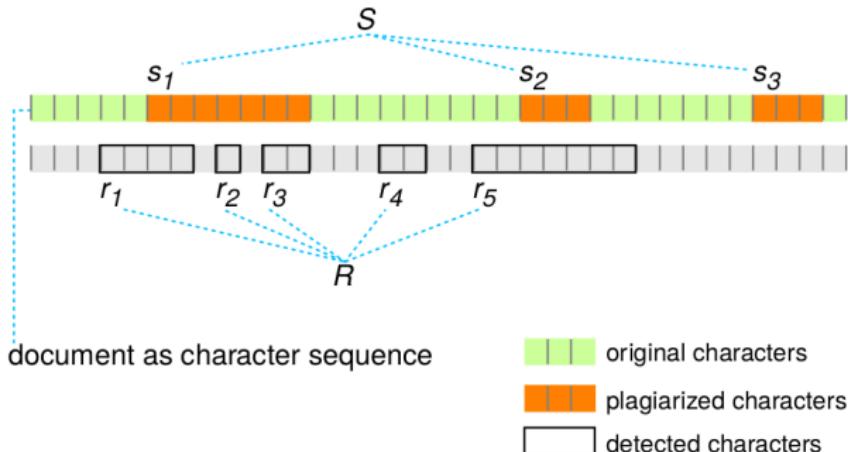
$$precision = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{retrieved\ documents\}|}$$

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

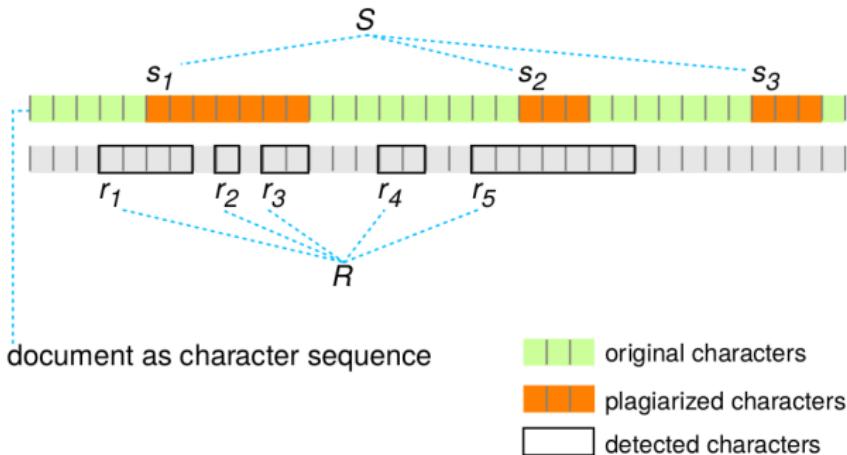
F-measure

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

PAN: Evaluation Measures - P and R

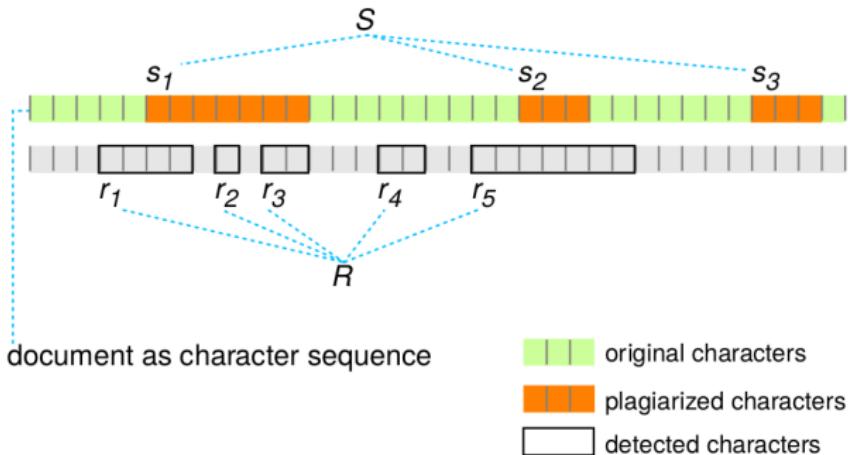


PAN: Evaluation Measures - P and R



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

PAN: Evaluation Measures - P and R

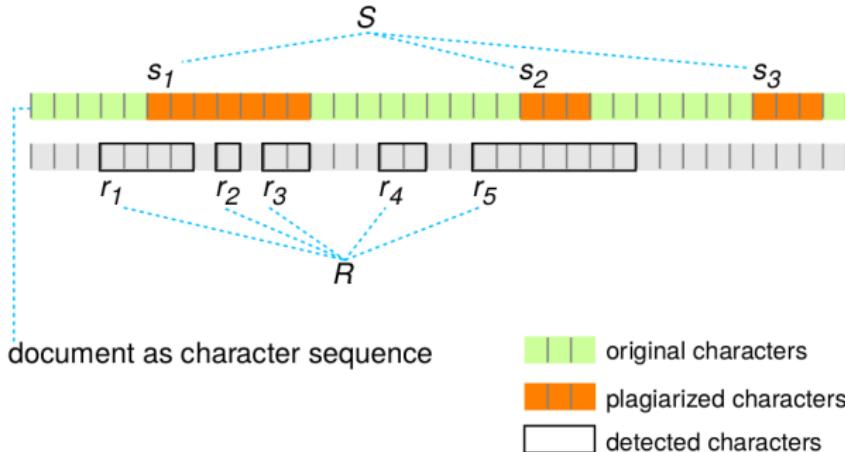


$$rec_{PDA}(S, R) = \frac{1}{|S|} \sum_{s \in S} \frac{|s \sqcap \bigcup_{r \in R} r|}{|s|}$$

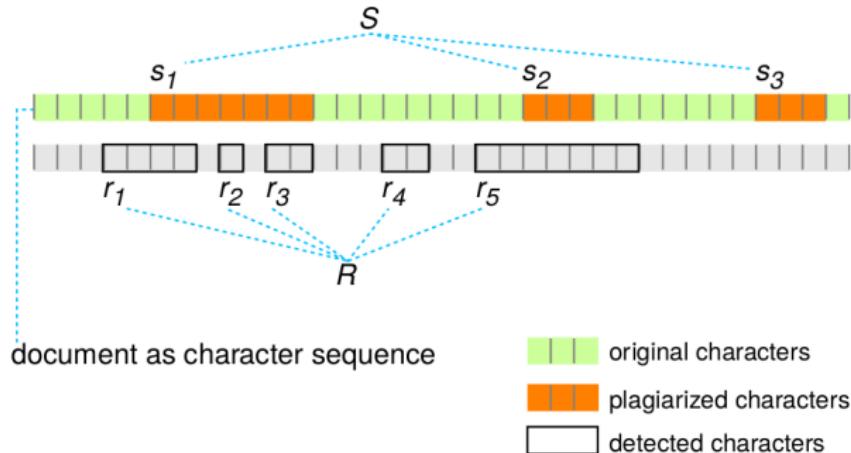
$$prec_{PDA}(S, R) = \frac{1}{|R|} \sum_{r \in R} \frac{|r \sqcap \bigcup_{s \in S} s|}{|r|}$$

(\sqcap computes the positionally overlapping characters)

PAN: Evaluation Measures - Granularity



PAN: Evaluation Measures - Granularity

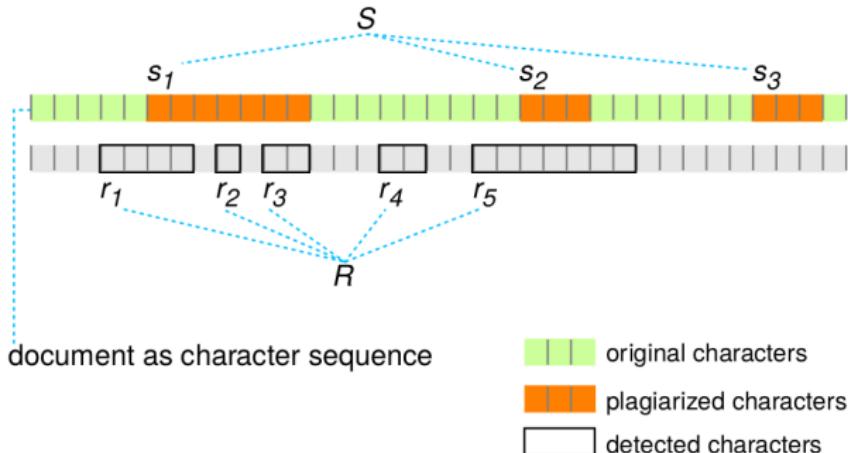


$$gran_{PDA}(S, R) = \frac{1}{|S_R|} \sum_{s \in S_R} |C_s| \quad \in [1, |R|]$$

$$C_s = \{r \mid r \in R \wedge s \cap r \neq \emptyset\}$$

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

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 prize, sponsored by Yahoo! Research.

PAN: 1st International Competition - Chronology

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

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).

PAN: 1st International Competition - Chronology

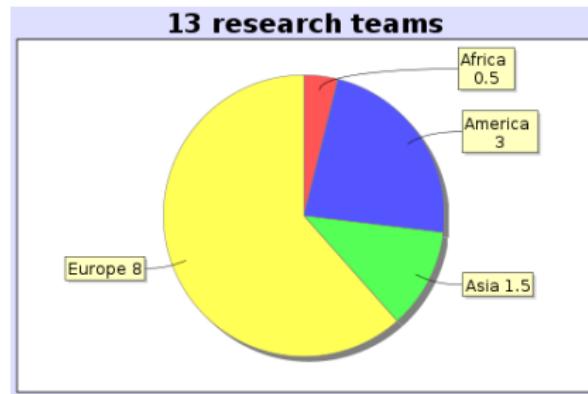
- March 2009 Participants were provided with the developing section of the corpus (with annotated cases).
- May 2009 Test corpus provided (without any annotation).
- June 2009 Participants submitted their detections to be evaluated.

PAN: 1st International Competition - 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

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

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

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>

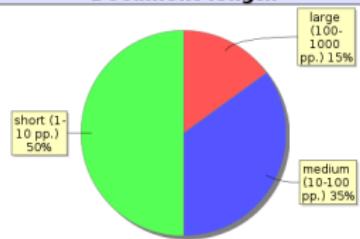
PAN-PC-10 Corpus

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

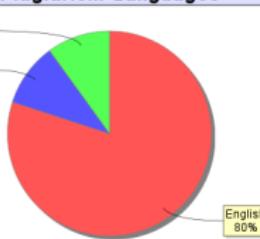
www.webis.de/research/corpora/pan-pc-10

PAN-PC-10 Corpus Parameters

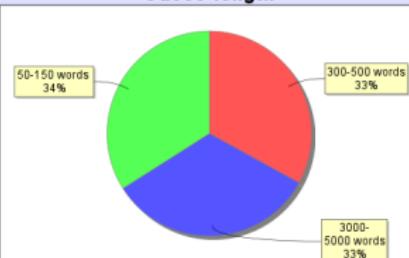
Document length



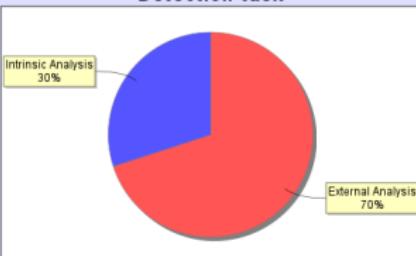
Plagiarism Languages



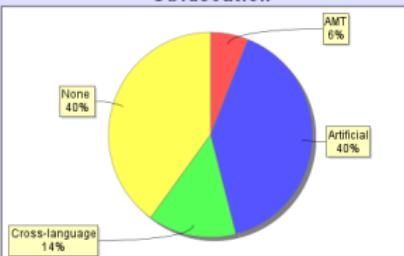
Cases length



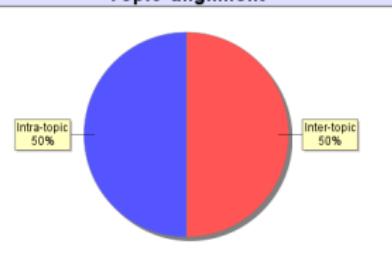
Detection task



Obfuscation



Topic alignment



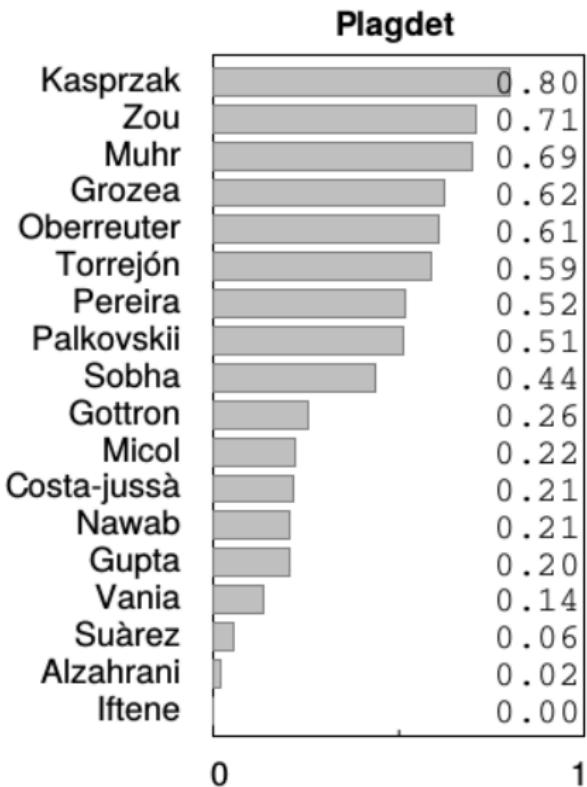
PAN: 2nd International Competition

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

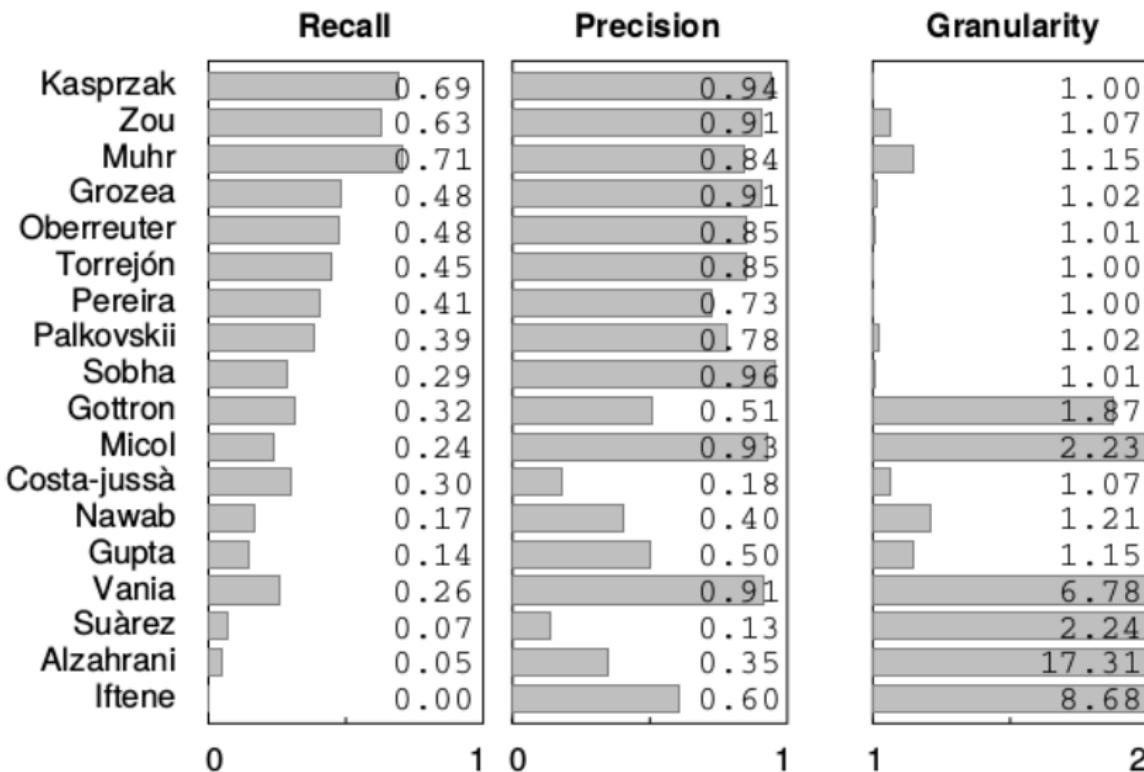
May 2010 Test corpus provided (brand new)

June 2010 Participants submitted their detections to be evaluated.

PAN: 2nd International Competition Results



PAN: 2nd International Competition Results



PAN: PAN-PC Corpora

- Synthetic cases of plagiarism
- Base docs.: Project Gutenberg
- Borrowed and source fragments clearly identified
- Cases for intrinsic and external detection included
- Large scale and size variety
- Languages: en (de and es)

[Potthast et al., 2010]

PAN: PAN-PC Corpora

- Synthetic cases of plagiarism
- Base docs.: Project Gutenberg
- Borrowed and source fragments clearly identified
- Cases for intrinsic and external detection included
- Large scale and size variety
- Languages: en (de and es)

[Potthast et al., 2010]

Documents and fragments selection



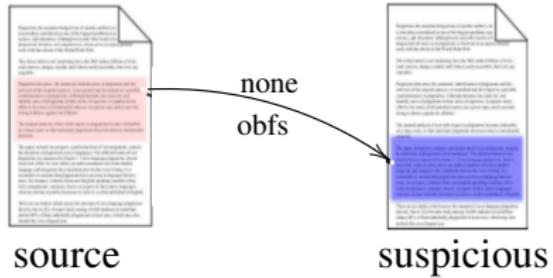
- * Preliminary clustering of the base documents in order to generate intra-topic plagiarism cases

PAN: PAN-PC Corpora

- Synthetic cases of plagiarism
- Base docs.: Project Gutenberg
- Borrowed and source fragments clearly identified
- Cases for intrinsic and external detection included
- Large scale and size variety
- Languages: en (de and es)

[Potthast et al., 2010]

Obfuscation (automatic)

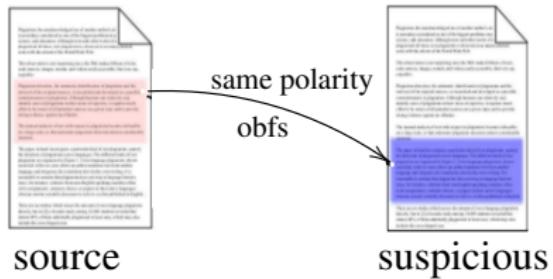


PAN: PAN-PC Corpora

- Synthetic cases of plagiarism
 - Base docs.: Project Gutenberg
 - Borrowed and source fragments clearly identified
 - Cases for intrinsic and external detection included
 - Large scale and size variety
 - Languages: en (de and es)

[Potthast et al., 2010]

Obfuscation (automatic)

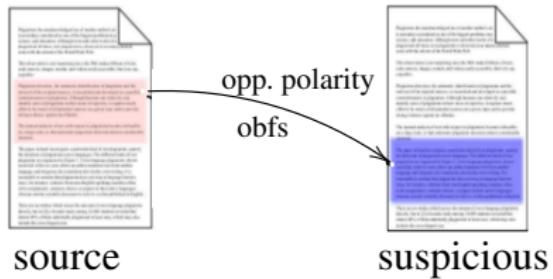


PAN: PAN-PC Corpora

- Synthetic cases of plagiarism
 - Base docs.: Project Gutenberg
 - Borrowed and source fragments clearly identified
 - Cases for intrinsic and external detection included
 - Large scale and size variety
 - Languages: en (de and es)

[Potthast et al., 2010]

Obfuscation (automatic)

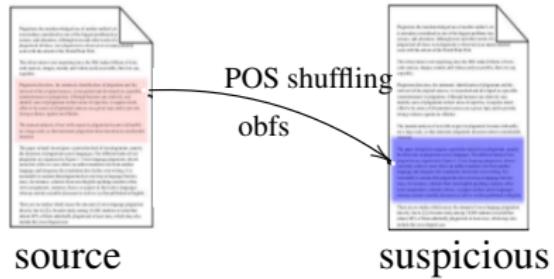


PAN: PAN-PC Corpora

- Synthetic cases of plagiarism
 - Base docs.: Project Gutenberg
 - Borrowed and source fragments clearly identified
 - Cases for intrinsic and external detection included
 - Large scale and size variety
 - Languages: en (de and es)

[Potthast et al., 2010]

Obfuscation (automatic)

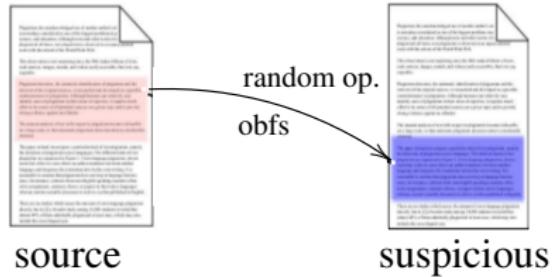


PAN: PAN-PC Corpora

- Synthetic cases of plagiarism
- Base docs.: Project Gutenberg
- Borrowed and source fragments clearly identified
- Cases for intrinsic and external detection included
- Large scale and size variety
- Languages: en (de and es)

[Potthast et al., 2010]

Obfuscation (automatic)

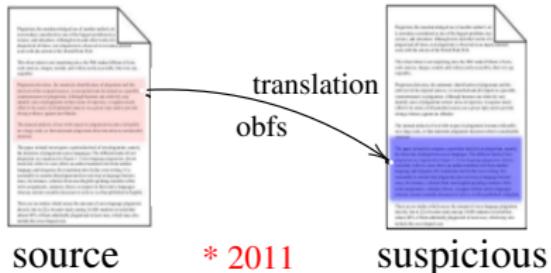


PAN: PAN-PC Corpora

- Synthetic cases of plagiarism
- Base docs.: Project Gutenberg
- Borrowed and source fragments clearly identified
- Cases for intrinsic and external detection included
- Large scale and size variety
- Languages: en (de and es)

[Potthast et al., 2010]

Obfuscation (automatic)



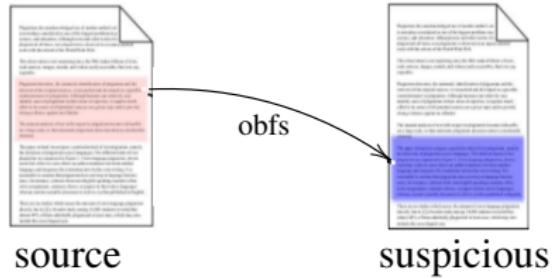
* The PAN-PC-11 included further obfuscated translated cases

PAN: PAN-PC Corpora

- Synthetic cases of plagiarism
- Base docs.: Project Gutenberg
- Borrowed and source fragments clearly identified
- Cases for intrinsic and external detection included
- Large scale and size variety
- Languages: en (de and es)

[Potthast et al., 2010]

Obfuscation (manual)



PAN: Impact of the Competition

The participants of the competitions run plagiarism detection systems in a real scenario:

www.theses.cz

[Kasprzak and Brandejs, 2010]
(best performance in 2010)

plagiarism-detector.com

[Palkovskii et al., 2011]
(regular participant)

www.docode.cl

[Oberreuter et al., 2011]
(best performance in 2011)

www.svop.sk/en/antiplag.aspx

[Grman and Ravas, 2011]
(second best performance in 2011)

PAN: Impact of the Competition

The participants of the competitions run plagiarism detection systems in a real scenario:

www.theses.cz

[Kasprzak and Brandejs, 2010]
(best performance in 2010)

plagiarism-detector.com

[Palkovskii et al., 2011]
(regular participant)

www.docode.cl

[Oberreuter et al., 2011]
(best performance in 2011)

www.svop.sk/en/antiplag.aspx

[Grman and Ravas, 2011]
(second best performance in 2011)

PAN has fostered research and motivated the development of systems that are used in academic (and other) scenarios.

PAN@CLEF: Lessons & Frontiers

- (sorted) word [3, . . . , 5]-grams and char. 16-grams are the best terms
 - cases of verbatim plagiarism are not challenging any longer
 - offset definition (fragment level detection) is a key factor to succeed
-

- cross-language and paraphrase plagiarism are far to be solved
- intrinsic detection remains an open issue
- Web-scale is necessary

PAN@CLEF: Lessons & Frontiers

- (sorted) word [3, . . . , 5]-grams and char. 16-grams are the best terms
 - cases of verbatim plagiarism are not challenging any longer
 - offset definition (fragment level detection) is a key factor to succeed
-

- cross-language and paraphrase plagiarism are far to be solved
- intrinsic detection remains an open issue
- Web-scale is necessary → PAN 2012

PAN@CLEF: Lessons & Frontiers

- (sorted) word [3, ..., 5]-grams and char. 16-grams are the best terms
 - cases of verbatim plagiarism are not challenging any longer
 - offset definition (fragment level detection) is a key factor to succeed
-

- cross-language and paraphrase plagiarism are far to be solved
- intrinsic detection remains an open issue
- Web-scale is necessary → PAN 2012

Outline

Introduction

Basic Concepts

Intrinsic Plagiarism Detection

External Plagiarism Detection

PAN Task on Plagiarism Detection

Plagiarism and Paraphrasing

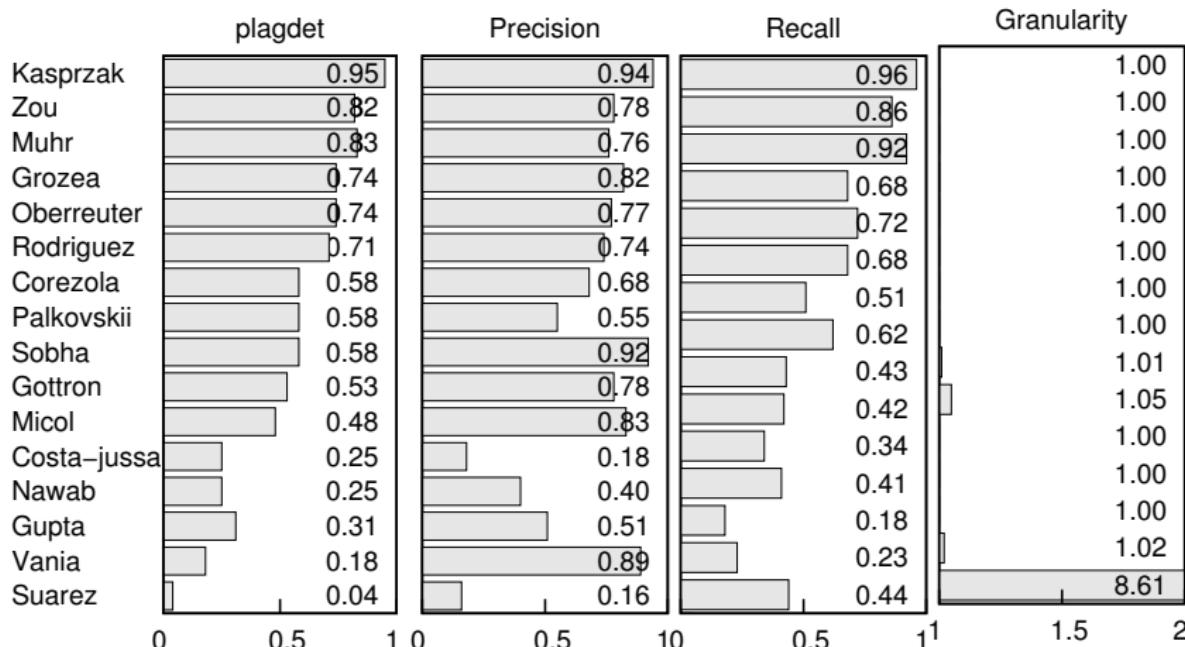
Cross-Language Plagiarism Detection

PAN Tasks @ CLEF

Detection of Plagiarism in Source Code

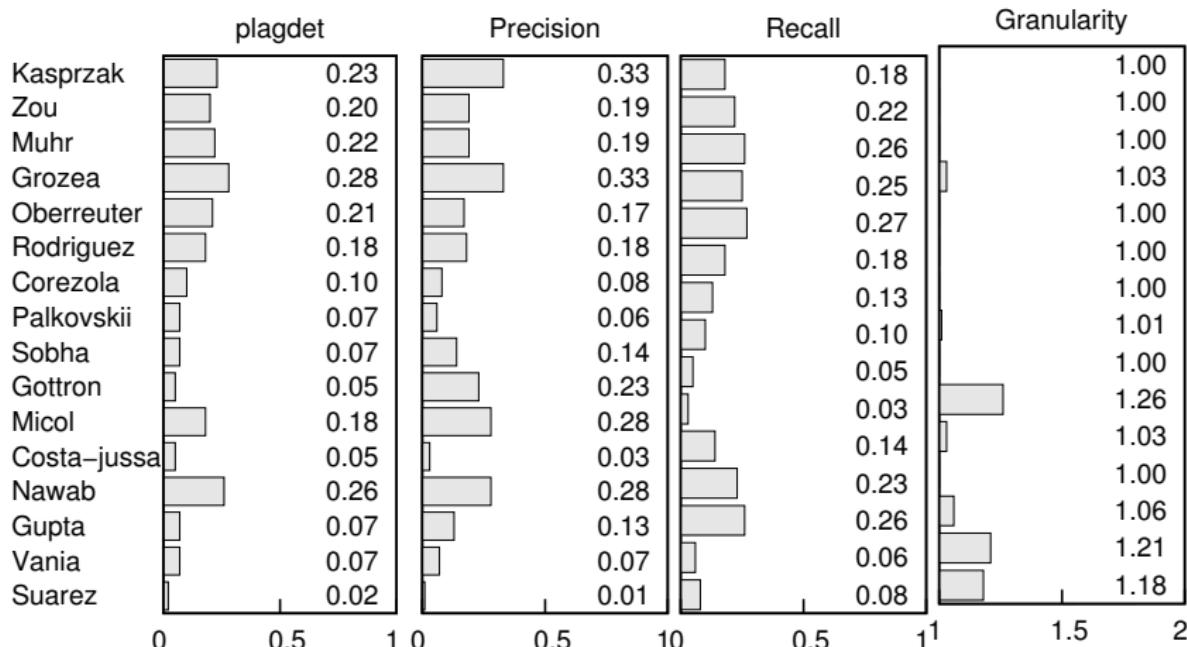
P4P: Detecting Paraphrase Plag (PAN 2010)

Verbatim



P4P: Detecting Paraphrase Plag (PAN 2010)

Manual paraphrase (Mechanical Turk)



P4P: Paraphrases and Plagiarism

Paraphrasing sameness of meaning between different wordings

(joint work with Universitat de Barcelona)

[Barrón-Cedeño et al., 2013b]

P4P: Paraphrases and Plagiarism

Paraphrasing sameness of meaning between different wordings

- Paraphrasing is the linguistic mechanism many plagiarism cases rely on

(joint work with Universitat de Barcelona)

[Barrón-Cedeño et al., 2013b]

P4P: Paraphrases and Plagiarism

Paraphrasing sameness of meaning between different wordings

- Paraphrasing is the linguistic mechanism many plagiarism cases rely on
- Little attention is paid to the paraphrasing-plagiarism relationship

(joint work with Universitat de Barcelona)

[Barrón-Cedeño et al., 2013b]

P4P: Paraphrases and Plagiarism

Paraphrasing sameness of meaning between different wordings

- Paraphrasing is the linguistic mechanism many plagiarism cases rely on
- Little attention is paid to the paraphrasing-plagiarism relationship
- Researchers on plagiarism detection are gradually turning to paraphrases analysis (although from a different perspective)
[Burrows et al., 2012]

(joint work with Universitat de Barcelona)

[Barrón-Cedeño et al., 2013b]

P4P: Paraphrases and Plagiarism

Paraphrasing sameness of meaning between different wordings

- Paraphrasing is the linguistic mechanism many plagiarism cases rely on
- Little attention is paid to the paraphrasing-plagiarism relationship
- Researchers on plagiarism detection are gradually turning to paraphrases analysis (although from a different perspective)
[Burrows et al., 2012]

Purpose Analysing the paraphrasing strategies applied during the text re-use process in order to determine what **paraphrasing types** make plagiarism harder to be uncovered

(joint work with Universitat de Barcelona)

[Barrón-Cedeño et al., 2013b]

P4P: Paraphrases Typology

① Morphology-based changes

- Inflectional changes
- Modal verb changes
- Derivational changes

② Lexicon-based changes

- Spelling and format changes
- Same polarity substitutions
- Synthetic/analytic substitutions
- Opposite polarity substitutions
- Inverse substitutions

③ Syntax-based changes

- Diathesis alternations
- Negation switching
- Ellipsis
- Coordination changes
- Subordination and nesting changes

④ Discourse-based changes

- Punctuation and format changes
- Direct/indirect style alternations
- Sentence modality changes

⑤ Miscellaneous changes

- Syntax/discourse structure changes
- Change of order
- Addition/deletion

⑥ Semantics-based changes

[Recasens and Vila, 2010, Vila et al., 2011]

P4P: Paraphrases Typology

① Morphology-based changes

- Inflectional changes
- Modal verb changes
- Derivational changes

② Lexicon-based changes

- Spelling and format changes
- Same polarity substitutions
- Synthetic/analytic substitutions
- Opposite polarity substitutions
- Inverse substitutions

③ Syntax-based changes

- Diathesis alternations
- Negation switching
- Ellipsis
- Coordination changes
- Subordination and nesting changes

④ Discourse-based changes

- Punctuation and format changes
- Direct/indirect style alternations
- Sentence modality changes

⑤ Miscellaneous changes

- Syntax/discourse structure changes
- Change of order
- Addition/deletion

⑥ Semantics-based changes

[Recasens and Vila, 2010, Vila et al., 2011]

[Clough et al., 2002]

[Clough and Gaizauskas, 2009]

P4P: Paraphrases Typology (examples)

① Morphology-based changes
Inflectional changes

② Lexicon-based changes
Same polarity substitutions

③ Syntax-based changes
Diathesis alternations

④ Discourse-based changes
Direct/indirect style altern

⑤ Miscellaneous changes
Addition/deletion

⑥ Semantics-based changes

- ① (a) You couldn't even follow the path of the **street**
(b) [...] the course of **streets** could be followed
- ② (a) **very little** vanilla
(b) **teaspoonful of** vanilla
- ③ (a) our attention **was drawn** by our guide to a little [...]
(b) the guide **drew** our attention to a gloomy little [...]
- ④ (a) The Great Spirit said that she is her
(b) “She is mine,” said the Great Spirit
- ⑤ (a) she took a hot flat-iron...
(b) **One day** she took a hot flat-iron
- ⑥ (a) Which added to the tropical appearance
(b) The scenery was altogether more tropical

P4P Paraphrases For Plagiarism Corpus

- Subsample of the simulated cases in the PAN-PC-10

P4P Paraphrases For Plagiarism Corpus

- Subsample of the simulated cases in the PAN-PC-10
- 847 plagiarism-source pairs

P4P Paraphrases For Plagiarism Corpus

- Subsample of the simulated cases in the PAN-PC-10
- 847 plagiarism-source pairs
- Cases shorter than 50 words

20 and 28 on average in other paraphrase corpora
[Barzilay and Lee, 2003, Dolan and Brockett, 2005]

P4P Paraphrases For Plagiarism Corpus

- Subsample of the simulated cases in the PAN-PC-10
- 847 plagiarism-source pairs
- Cases shorter than 50 words
- Annotation process carried out at the Universitat de Barcelona

P4P: Corpus Statistics

	f_{abs}	f_{rel}
Morphology-based changes	631	0.057
Inflectional changes	254	0.023
Modal verb changes	116	0.010
Derivational changes	261	0.024
Lexicon-based changes	6,264	0.564
Spelling and format ch.	436	0.039
Same polarity subst.	5,056	0.456
Synthetic/analytic subst.	658	0.059
Opposite polarity subst.	65	0.006
Inverse substitutions	33	0.003
Syntax-based changes	1,045	0.094
Diathesis alternations	128	0.012
Negation switching	33	0.003
Ellipsis	83	0.007
Coordination changes	188	0.017
Subord. and nesting ch.	484	0.044

	f_{abs}	f_{rel}
Discourse-based changes	501	0.045
Punctuation and format ch.	430	0.039
Direct/indirect style altern.	36	0.003
Sentence modality changes	35	0.003
Miscellaneous changes	2,331	0.210
Syntax/discourse structure ch.	304	0.027
Change of order	556	0.050
Addition/deletion	1471	0.132
Semantics-based changes	335	0.030
Others	136	0.012
Identical	101	0.009
Non paraphrases	35	0.003

P4P: Corpus Statistics

	f_{abs}	f_{rel}
Morphology-based changes	631	0.057
Inflectional changes	254	0.023
Modal verb changes	116	0.010
Derivational changes	261	0.024
Lexicon-based changes	6,264	0.564
Spelling and format ch.	436	0.039
Same polarity subst.	5,056	0.456
Synthetic/analytic subst.	658	0.059
Opposite polarity subst.	65	0.006
Inverse substitutions	33	0.003
Syntax-based changes	1,045	0.094
Diathesis alternations	128	0.012
Negation switching	33	0.003
Ellipsis	83	0.007
Coordination changes	188	0.017
Subord. and nesting ch.	484	0.044

	f_{abs}	f_{rel}
Discourse-based changes	501	0.045
Punctuation and format ch.	430	0.039
Direct/indirect style altern.	36	0.003
Sentence modality changes	35	0.003
Miscellaneous changes	2,331	0.210
Syntax/discourse structure ch.	304	0.027
Change of order	556	0.050
Addition/deletion	1471	0.132
Semantics-based changes	335	0.030
Others	136	0.012
Identical	101	0.009
Non paraphrases	35	0.003

P4P: Corpus Statistics

	f_{abs}	f_{rel}
Morphology-based changes	631	0.057
Inflectional changes	254	0.023
Modal verb changes	116	0.010
Derivational changes	261	0.024
Lexicon-based changes	6,264	0.564
Spelling and format ch.	436	0.039
Same polarity subst.	5,056	0.456
Synthetic/analytic subst.	658	0.059
Opposite polarity subst.	65	0.006
Inverse substitutions	33	0.003
Syntax-based changes	1,045	0.094
Diathesis alternations	128	0.012
Negation switching	33	0.003
Ellipsis	83	0.007
Coordination changes	188	0.017
Subord. and nesting ch.	484	0.044

	f_{abs}	f_{rel}
Discourse-based changes	501	0.045
Punctuation and format ch.	430	0.039
Direct/indirect style altern.	36	0.003
Sentence modality changes	35	0.003
Miscellaneous changes	2,331	0.210
Syntax/discourse structure ch.	304	0.027
Change of order	556	0.050
Addition/deletion	1471	0.132
Semantics-based changes	335	0.030
Others	136	0.012
Identical	101	0.009
Non paraphrases	35	0.003

P4P: Corpus Statistics

	f_{abs}	f_{rel}
Morphology-based changes	631	0.057
Inflectional changes	254	0.023
Modal verb changes	116	0.010
Derivational changes	261	0.024
Lexicon-based changes	6,264	0.564
Spelling and format ch.	436	0.039
Same polarity subst.	5,056	0.456
Synthetic/analytic subst.	658	0.059
Opposite polarity subst.	65	0.006
Inverse substitutions	33	0.003
Syntax-based changes	1,045	0.094
Diathesis alternations	128	0.012
Negation switching	33	0.003
Ellipsis	83	0.007
Coordination changes	188	0.017
Subord. and nesting ch.	484	0.044

	f_{abs}	f_{rel}
Discourse-based changes	501	0.045
Punctuation and format ch.	430	0.039
Direct/indirect style altern.	36	0.003
Sentence modality changes	35	0.003
Miscellaneous changes	2,331	0.210
Syntax/discourse structure ch.	304	0.027
Change of order	556	0.050
Addition/deletion	1471	0.132
Semantics-based changes	335	0.030
Others	136	0.012
Identical	101	0.009
Non paraphrases	35	0.003

in general $\text{len}_{src} > \text{len}_{plg}$. People tend to:
summarise, shorten expressions, or delete some fragments

P4P: Structuring the Corpus

Organisation of the cases according to the occurring paraphrase phenomena

- k -means clustering method [MacQueen, 1967]
- feature vectors of 21 dimensions (no same polarity substitutions)
- weighting: relative frequency of the type in the plagiarism case

P4P: Structuring the Corpus

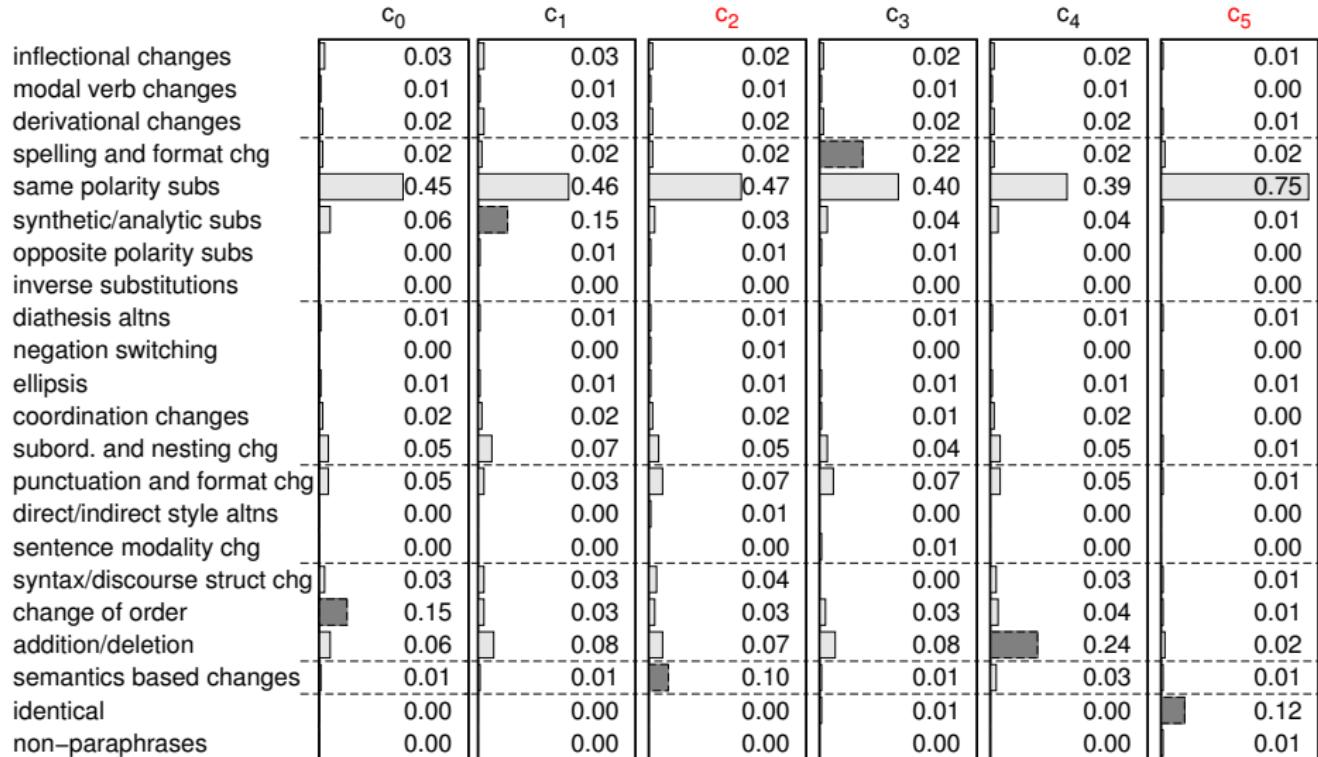
Organisation of the cases according to the occurring paraphrase phenomena

- k -means clustering method [MacQueen, 1967]
- feature vectors of 21 dimensions (no same polarity substitutions)
- weighting: relative frequency of the type in the plagiarism case
- $k = 6$ after applying the Elbow method

P4P: Structuring the Corpus (insights)

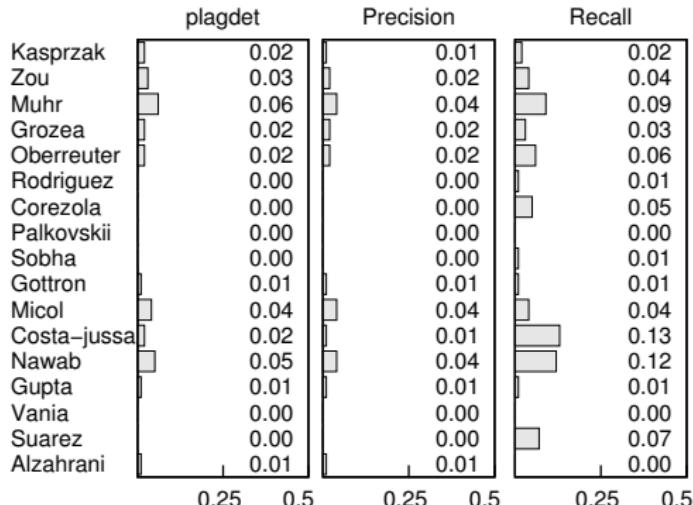
	c_0	c_1	c_2
inflectional changes	0.03	0.03	0.02
modal verb changes	0.01	0.01	0.01
derivational changes	0.02	0.03	0.02
spelling and format chg	0.02	0.02	0.02
same polarity subs	0.45	0.46	0.47
synthetic/analytic subs	0.06	0.15	0.03
opposite polarity subs	0.00	0.01	0.01
inverse substitutions	0.00	0.00	0.00
diathesis altns	0.01	0.01	0.01
negation switching	0.00	0.00	0.01
ellipsis	0.01	0.01	0.01
coordination changes	0.02	0.02	0.02
subord. and nesting chg	0.05	0.07	0.05
punctuation and format chg	0.05	0.03	0.07
direct/indirect style altns	0.00	0.00	0.01
sentence modality chg	0.00	0.00	0.00
syntax/discourse struct chg	0.03	0.03	0.04
change of order	0.15	0.03	0.03
addition/deletion	0.06	0.08	0.07
semantics based changes	0.01	0.01	0.10
identical	0.00	0.00	0.00
non-paraphrases	0.00	0.00	0.00

P4P: Structuring the Corpus (insights)



P4P: PAN 2010 Competitors P4P Performance

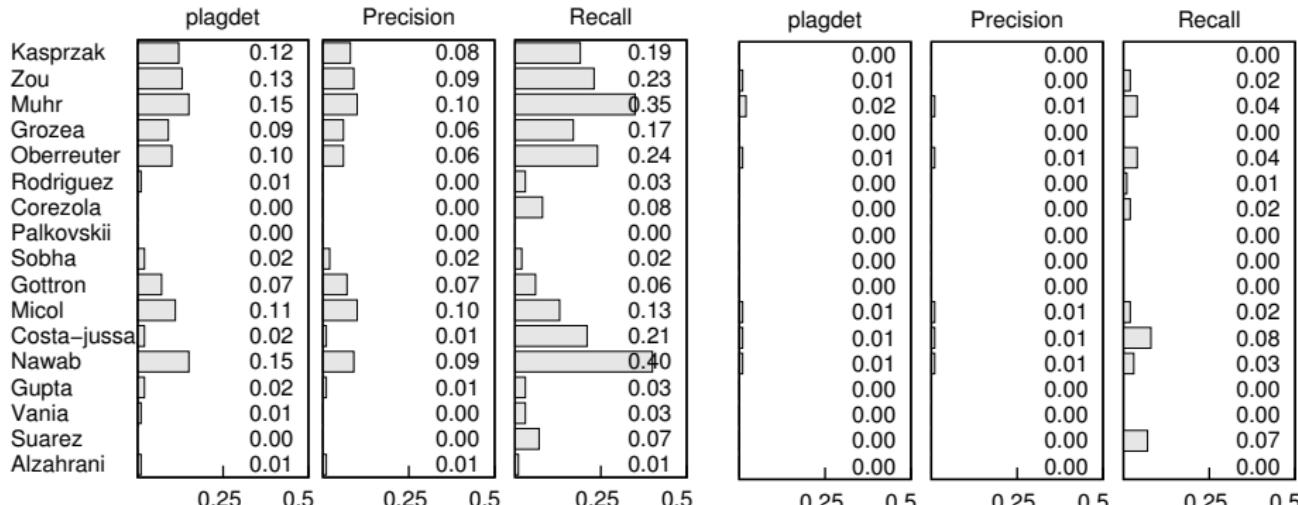
overall



P4P: PAN 2010 Competitors P4P Performance

c_5

c_2



P4P: Discussion

- More linguistic and quantitative complexity implies worse performance of the systems

P4P: Discussion

- More linguistic and quantitative complexity implies worse performance of the systems
- The best performing models do not better handle paraphrases. They uncover “surrounding” verbatim copied fragments, such as in c_5 (GST by [Nawab et al., 2010] and dotplot by [R. Costa-jussà et al., 2010])

P4P: Discussion

- More linguistic and quantitative complexity implies worse performance of the systems
- The best performing models do not better handle paraphrases. They uncover “surrounding” verbatim copied fragments, such as in c_5 (GST by [Nawab et al., 2010] and dotplot by [R. Costa-jussà et al., 2010])
- Still they pay attention to “typical” paraphrasing: case folding, stopword removal, stemming [R. Costa-jussà et al., 2010]

P4P: Discussion

- More linguistic and quantitative complexity implies worse performance of the systems
- The best performing models do not better handle paraphrases. They uncover “surrounding” verbatim copied fragments, such as in c_5 (GST by [Nawab et al., 2010] and dotplot by [R. Costa-jussà et al., 2010])
- Still they pay attention to “typical” paraphrasing: case folding, stopword removal, stemming [R. Costa-jussà et al., 2010]
- Lexical substitutions are the paraphrase mechanisms used the most
→Wordnet? (Babelnet or Eurovoc for CL)

P4P: Discussion

- More linguistic and quantitative complexity implies worse performance of the systems
- The best performing models do not better handle paraphrases. They uncover “surrounding” verbatim copied fragments, such as in c_5 (GST by [Nawab et al., 2010] and dotplot by [R. Costa-jussà et al., 2010])
- Still they pay attention to “typical” paraphrasing: case folding, stopword removal, stemming [R. Costa-jussà et al., 2010]
- Lexical substitutions are the paraphrase mechanisms used the most →Wordnet? (Babelnet or Eurovoc for CL)
- The paraphrase mechanisms tend to produce a summarised version of the re-used text → length model?

Outline

Introduction

Basic Concepts

Intrinsic Plagiarism Detection

External Plagiarism Detection

PAN Task on Plagiarism Detection

Plagiarism and Paraphrasing

Cross-Language Plagiarism Detection

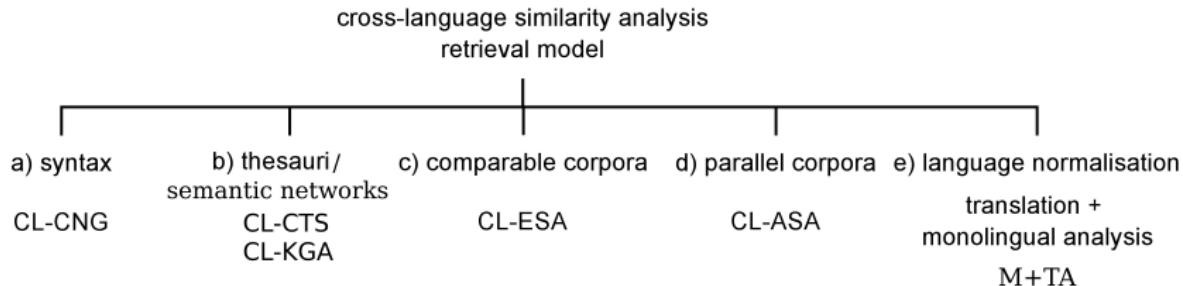
PAN Tasks @ CLEF

Detection of Plagiarism in Source Code

CL Plagiarism Detection

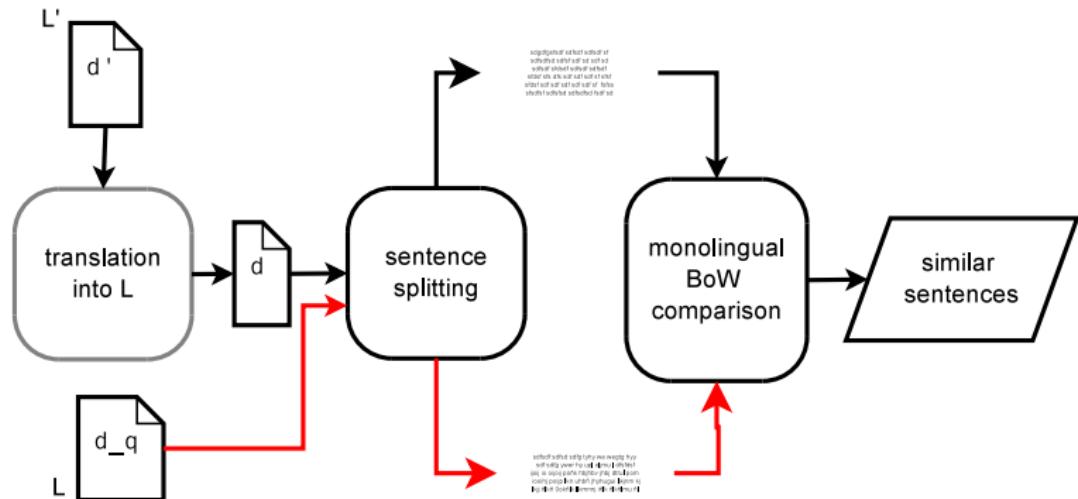
- Researchers are still forging the state of art in CL plagiarism detection
- The most of the methods are based on previously proposed models for CLIR

CL: Methods Overview

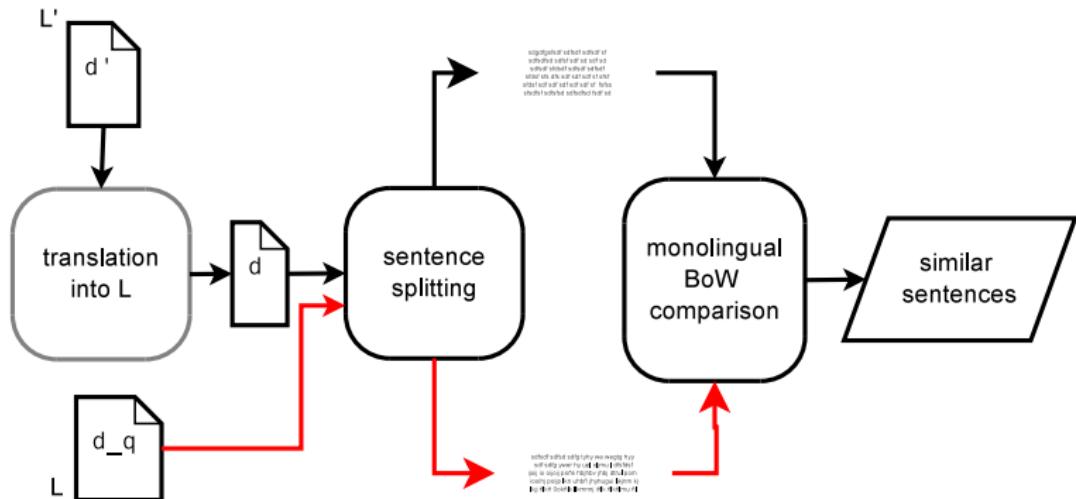


(before 2008 no technology for cross-language plagiarism detection had been developed) [Barrón-Cedeño et al., 2008, Ceska et al., 2008, Lee et al., 2008, Pinto et al., 2009a, Potthast et al., 2008, Gupta et al., 2012, Franco-Salvador et al., 2013a, Barrón-Cedeño et al., 2013a]

CL: Translation + Monolingual Analysis



CL: Translation + Monolingual Analysis



The translation can be carried out on the basis of:

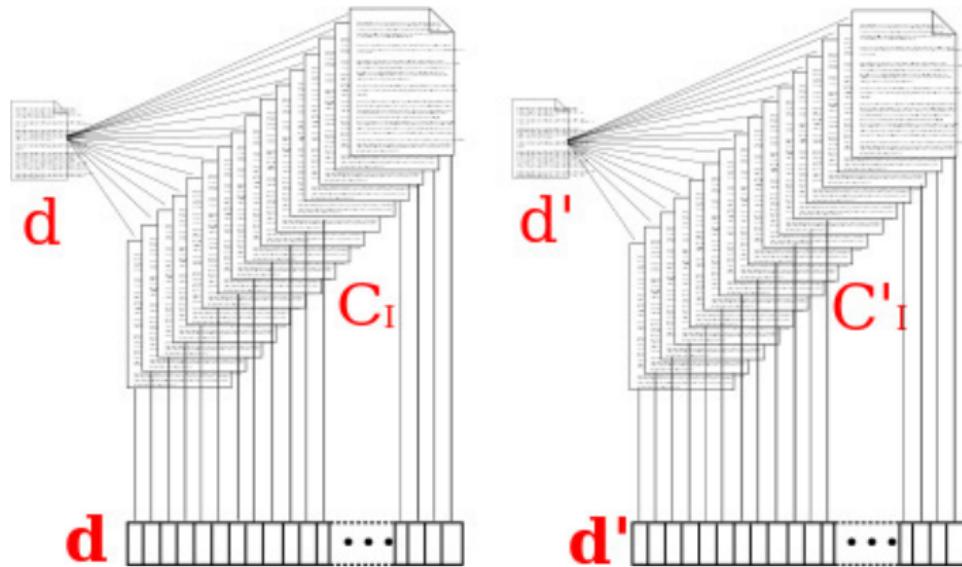
- Commercial MT systems (such as Google and Babelfish)
- Giza++, Moses, SRILM
[Och and Ney, 2003, Koehn et al., 2007, Stolcke, 2002]
- Considering multiple translations per word [Muhr et al., 2010]

CL-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-ESA



[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 the basic principles of statistical MT

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

Bayes' rule for statistical Machine Translation:

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

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

[Brown et al., 1993]

translation model

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

- $w(d \mid d')$ 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 $\epsilon = 0.1$.

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

Adapted translation model

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

- $w(d \mid d')$ 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 $\epsilon = 0.1$.

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

Length Model

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

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

Length Model

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

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

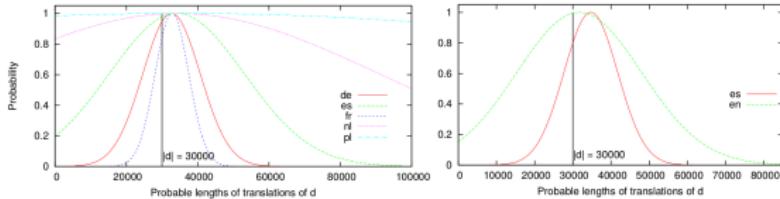
	en-de	en-es	en-fr	en-nl	en-pl	en-eu	es-eu
μ	1.089	1.138	1.093	1.143	1.216	1.0560	1.1569
σ	0.268	0.631	0.157	1.885	6.399	0.5452	0.2351

Length Model

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

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

	en-de	en-es	en-fr	en-nl	en-pl	en-eu	es-eu
μ	1.089	1.138	1.093	1.143	1.216	1.0560	1.1569
σ	0.268	0.631	0.157	1.885	6.399	0.5452	0.2351



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

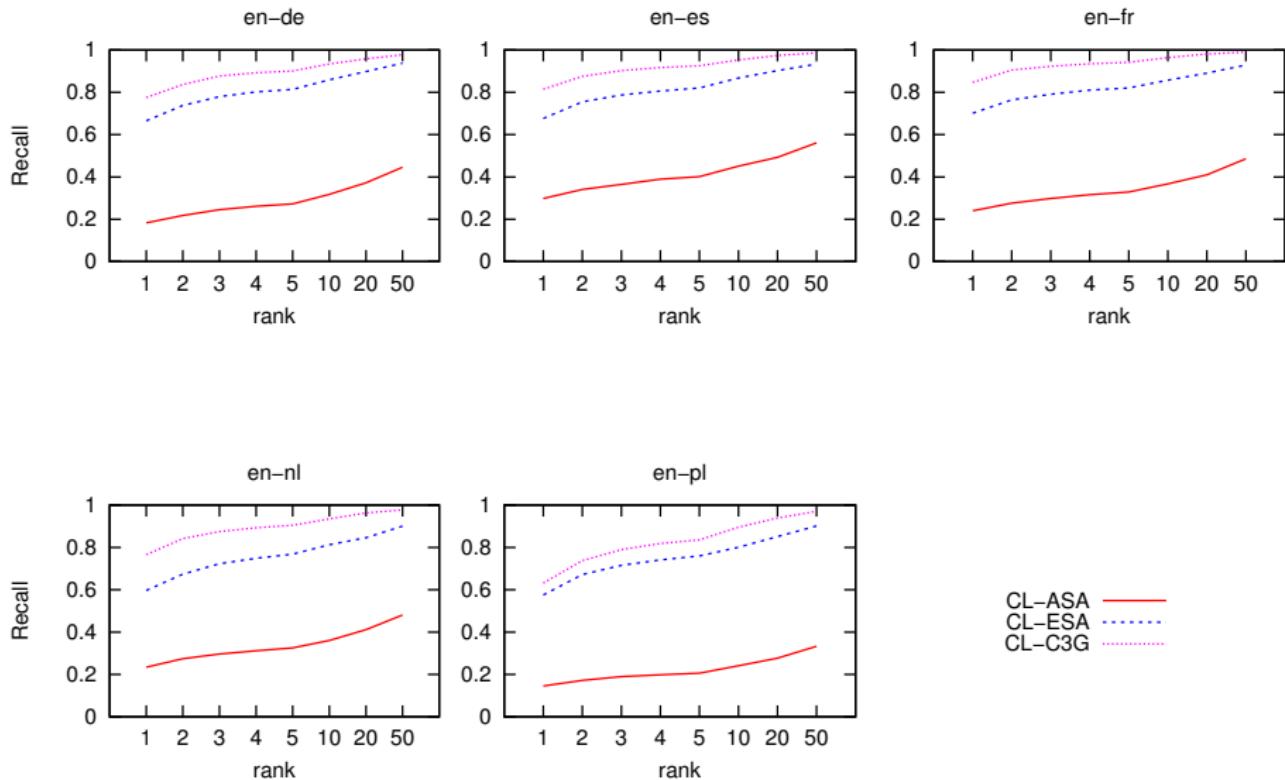
CL: Character n -grams

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

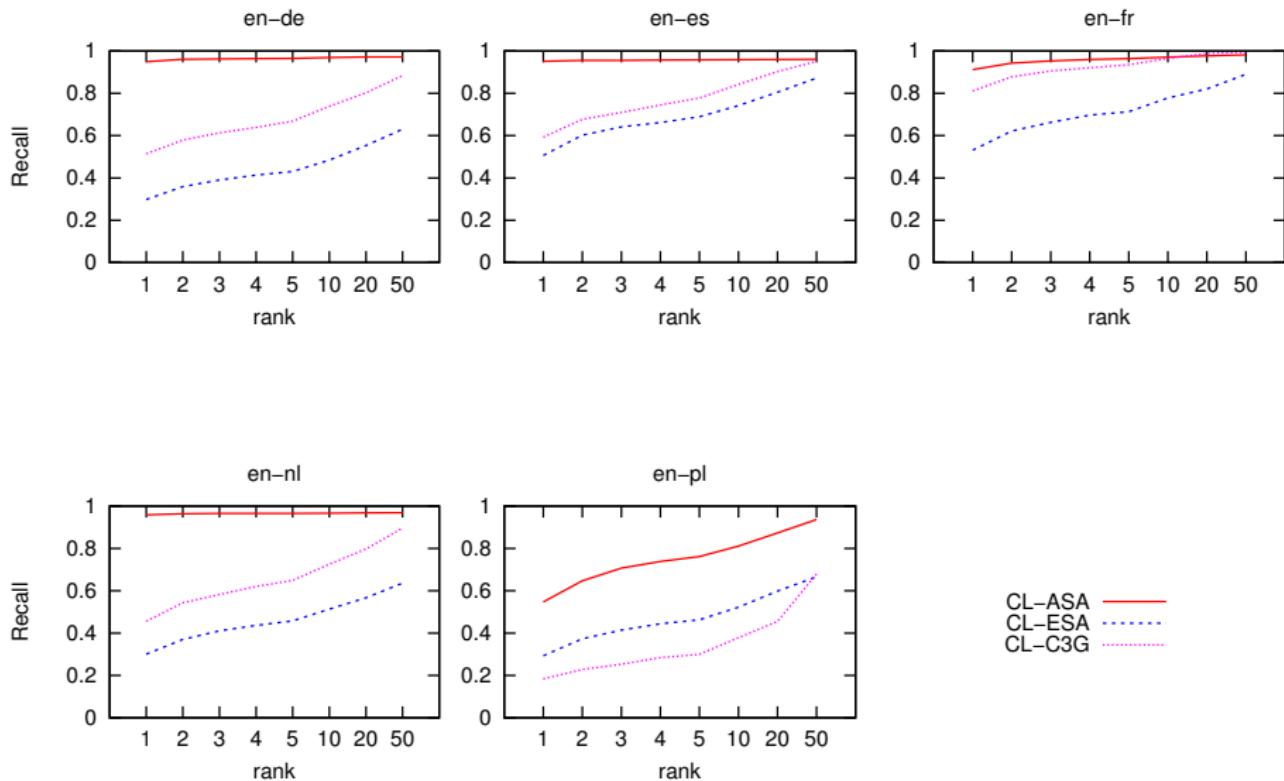
- $\Sigma = \{a, \dots, z, 0, \dots, 9\}$,
- $n = 3$
- $tfidf$ -weighting
- Cosine similarity

[Mcnamee and Mayfield, 2004]

CL: Cross-Language Ranking (Wikipedia)



CL: Cross-language ranking (JRC-Acquis)



CL: And for less resourced 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.

[Wikipedia, 2010b]

CL: And for less resourced 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]

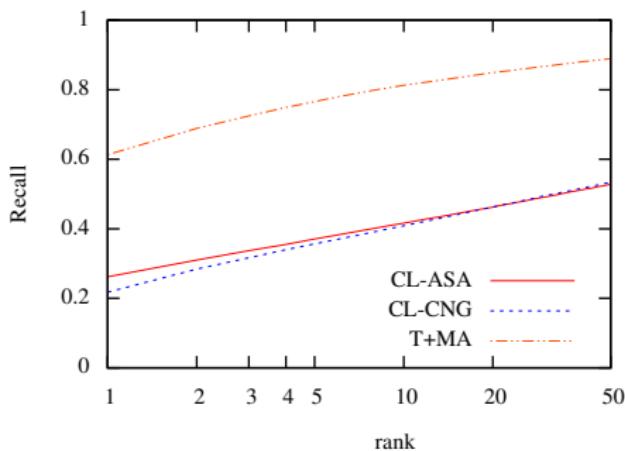
CL: Less Resourced Languages

Framework

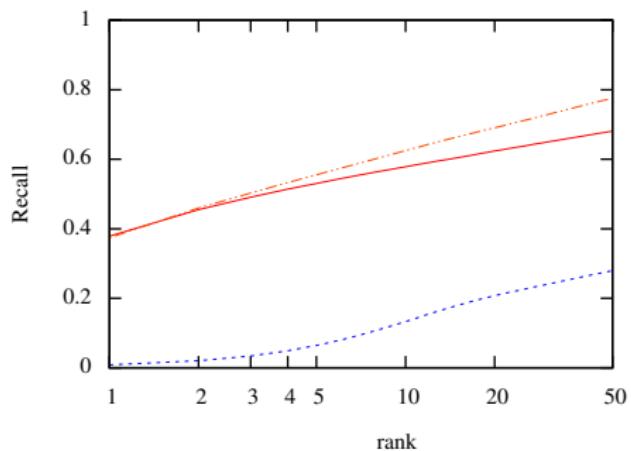
- Two parallel corpora:
 - software a translation memory (en-eu)
 - consumer extracts from a multilingual magazine (es-eu)
- The entire corpus is a “big” document
- We perform sentence level similarity estimation

(corpora provided by Elhuyar Fundazioa and Consumer)

CL: Less Resourced Languages



(a) es-eu



(b) en-eu

[Barrón-Cedeño et al., 2010]

EUROVOC Thesaurus-based

- Thesaurus catalogued manually
- Available in the 18 EU languages

[Pouliquen et al., 2003]

CL: Thesaurus based

EUROVOC Thesaurus-based

- Thesaurus catalogued manually
- Available in the 18 EU languages

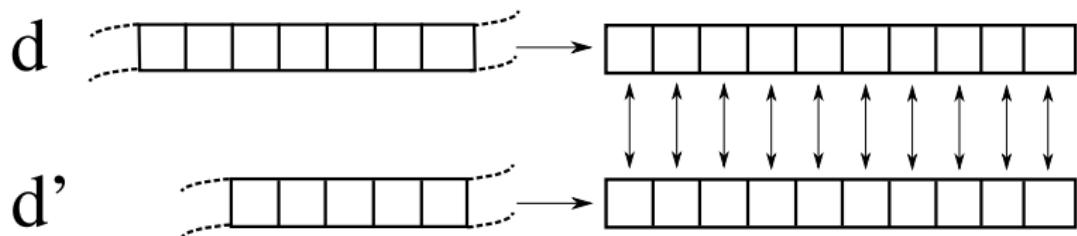
Example “transport of dangerous goods” lemmas

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-Conceptual Thesaurus based Similarity

- Represent documents as a vector of concepts
- Concept assignment is the least trivial part
- **Challenge:** Exploit a domain specific CT for all the corpora
- Assignment of concepts according to their verbatim occurrence in the document gives very bad results [Pouliquen et al., 2006]
- Assign a concept to a document if it “triggers the concept”

[Gupta et al., 2012]

Method contd.

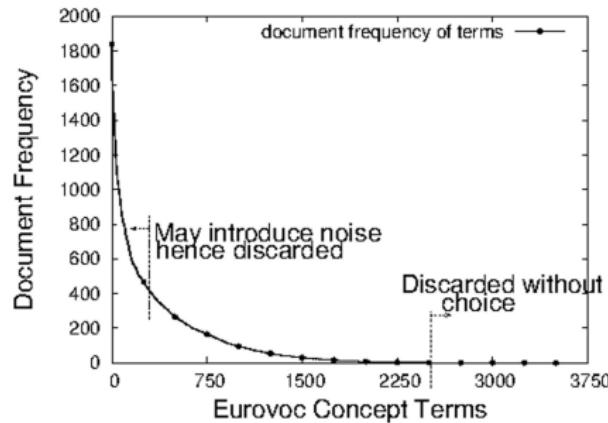
- **Heuristic:** The terms together are highly domain dependent but alone are domain independent.
- For example, “community” and “trade” compared to “community trade”

Concept Assignment

- Sum of the term frequencies (TF) of the terms in the concept in the Doc
- Stopword removal + stemming
- Filter the terms based on the discriminative power in the corpora

Method contd.

- All the concepts do not help in similarity estimation - Hence **Reduced Concepts (RC)**
 - Reduces the comparison vocabulary drastically
 - Domain independent threshold $0 < df(t) < \beta$
 - Automatic domain adaptation (**Football** in “Sports” and “Society and Culture”)



Method contd.

- **Concern** - The concepts are limited and are common across even slightly relevant documents
- To overcome the limitation of conceptual similarity estimation, we use Named Entities in similarity too
- n-gram similarity of NEs - **simplest method**
- NEs act as discriminative features - **e.g. Wikipedia page of Rome vs. Madrid**

Method contd.

- Sometimes high similar documents are parallel and the task is to find the parallel document for the given document
- A pattern in length is noticed for parallel documents across languages [Pouliquen et al., 2006]
- we use the same “length penalty”

$$\text{len}(\text{parallel}(d_q)) = f(\mu, \sigma, \text{len}(d_q))$$

Method contd.

- The similarity function

$$\omega(q, d) = \frac{\alpha}{2} * \left(\frac{\vec{c}_q \cdot \vec{c}_d}{|q||d|} + \ell(q, d) \right) + (1 - \alpha) * \zeta(q, d)$$

Conceptual Component NE Component

Conceptual Similarity Length Penalty

Compared with

- ① Cross-language Alignment based Similarity Analysis
(CL-ASA) [Barrón-Cedeño et al., 2008, Pinto et al., 2009b]
- ② Cross-language Character n-grams
(CL-CNG) [Mcnamee and Mayfield, 2004]

Datasets

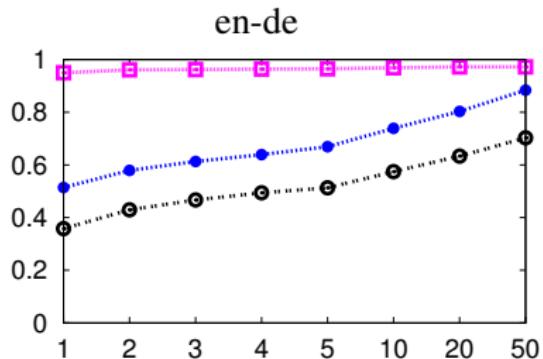
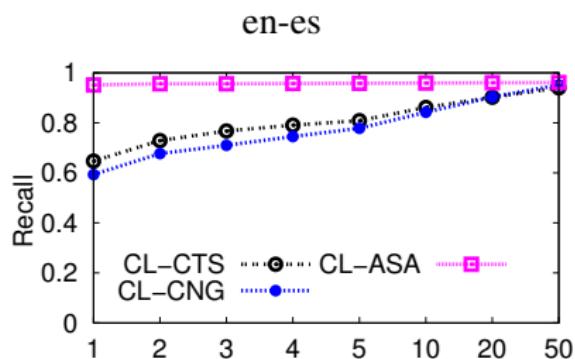
- JRC-Acquis (JRC)
 - Nature: related to European Commission activities
 - Size: 10,000 in each language
 - Type: Parallel
- PAN-PC-2011 (PAN)
 - Nature: Project Gutenberg (artificially created cross-language plagiarism cases)
 - Size: 2920 (en-es) and 2222 (en-de)
 - Type: Noisy parallel
- Wikipedia (Wiki)
 - Nature: General Wikipedia pages
 - Size: 10000 in each language
 - Type: Comparable

Datasets contd..

- Vocabulary shared by Eurovoc and JRC is higher than that of Eurovoc and PAN or Wiki.

CL-CTS

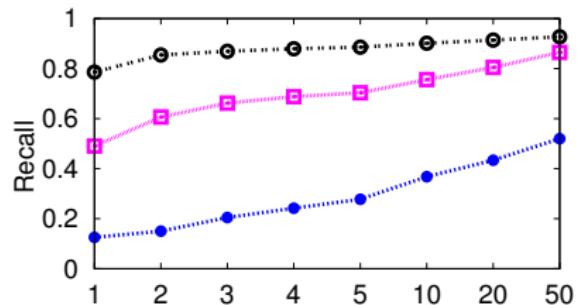
Results : JRC



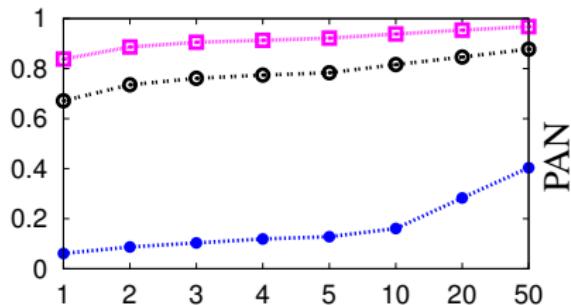
CL-CTS

Results : PAN

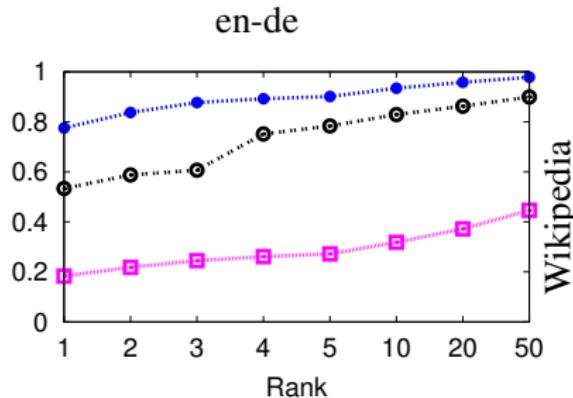
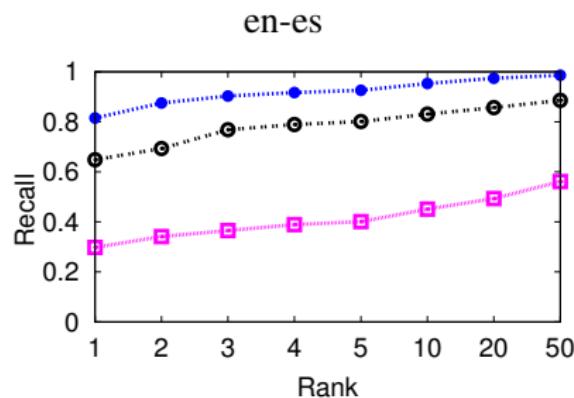
en-es



en-de



Results : Wiki



Wikipedia

Analysis

- Performance of CL-CTS with reduced concepts is much higher compared to inclusion of all concepts
 - R@1 0.02 → 0.58 (JRC en-es)
- Inclusion of NE component usually improves the performance except JRC - Interesting!
- CL-ASA and CL-CNG exhibit very corpus dependent performance.
- German stays more difficult compared to Spanish (compounding of the words needs better care)

Analysis: Further characterizing the corpora

- JRC
 - Parallel corpus
 - high amount of NEs
 - NEs are mostly of type ORG and LOC which appear quite identically in many documents
- PAN
 - Cross-language plagiarism cases artificially generated using SMT and/or manual correction - Noisy Parallel
 - documents are related to literature - contains far more natural language terms compared to NEs
 - NEs are mostly of type PERSON and is much diverse across documents

Analysis contd.

- Wiki
 - Generic documents - **comparable**
 - Lots of NEs, but diverse
- We investigated the distribution of NEs among corpora

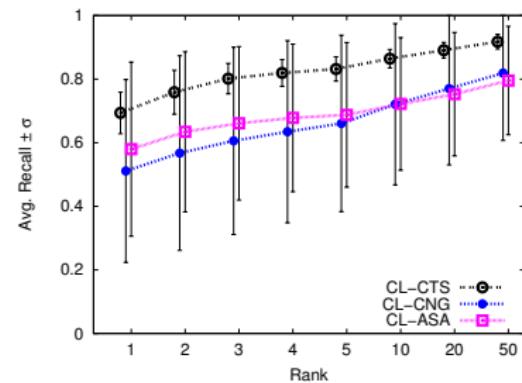
Corpus	Person	Location	Organisation	Total
JRC	1.8 %	2.3 %	8.7 %	12.9 %
PAN	1.8 %	1.7 %	1.9 %	5.4 %
Wiki	4.7 %	3.7 %	5.5 %	14.0 %

Observations

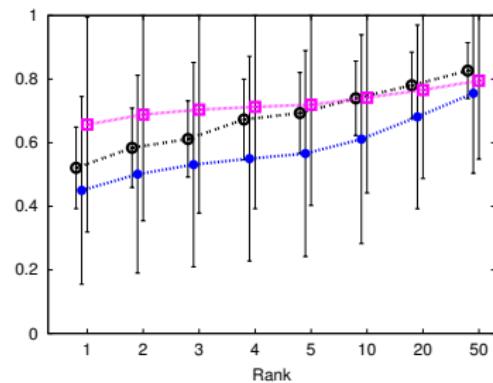
- CL-ASA performs better on the JRC and very poor on the Wiki
 - better results on nearly parallel data
- CL-CNG performs better on the Wiki and very poor on the PAN
 - better performance on the NE dominated corpora
- CL-CTS exhibits very stable performance across the corpora

Analysis : Average performance and standard deviation

en-es



en-de



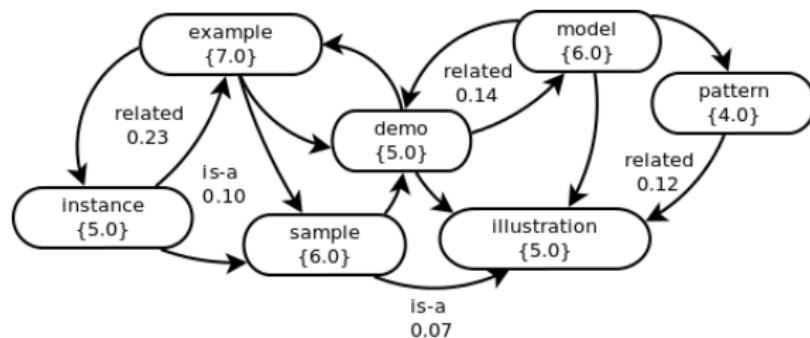
Remarks : CL-CTS

- Outperforms
 - char n-gram based model on linguistic corpus (PAN)
 - machine translation based model on comparable corpus (Wiki)
- Achieves a stable performance across the domains using a domain specific thesaurus
- Useful when
 - the nature of data is unknown OR
 - dealing with a heterogeneous data
- Uses reduced concepts and NEs → very compact inverted index and low computational cost

Knowledge graphs

A knowledge graph is a weighted and labeled graph that expands and relates the original concepts present in a set of words.

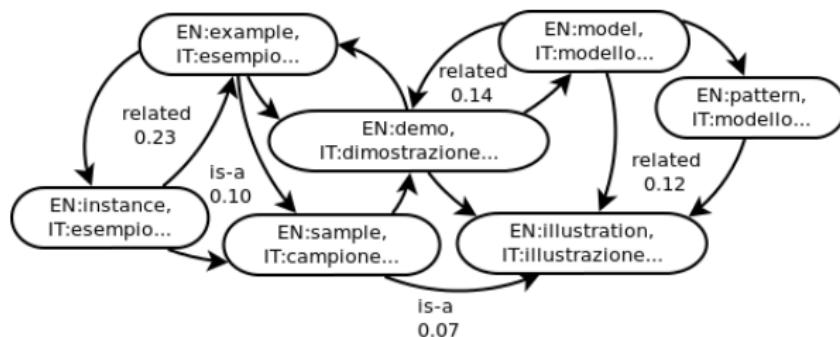
Concepts: {example, model}



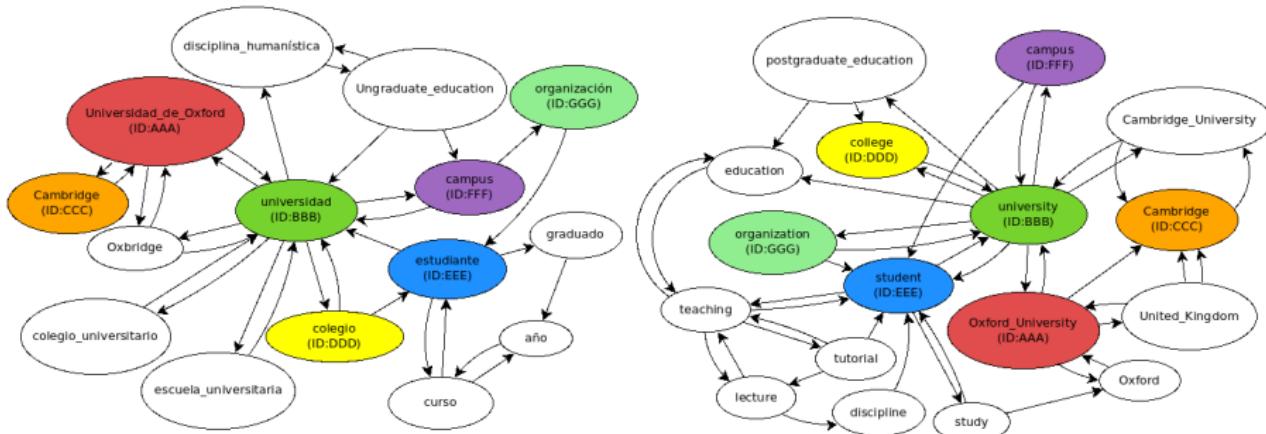
Knowledge graphs

A knowledge graph is a weighted and labeled graph that expands and relates the original concepts present in a set of words.

Concepts: {example, model}



Knowledge graphs



BabelNet

Knowledge graphs are built using the multilingual semantic network BabelNet¹ [Navigli and Ponzetto, 2012]:

- It consists of a labeled directed graph where nodes represent multilingual concepts and named entities, and edges express semantic relations between them.
- BabelNet 2.5 covers 50 languages.
- It integrates:
 - WordNet
 - Open Multilingual WordNet
 - Wikipedia
 - OmegaWiki
 - Wiktionary
 - Wikidata

¹babelnet.org

Cross-language plagiarism detection

ES: “La huelga comenzó oficialmente el 29 de mayo, y el 1 de junio los fabricantes se reunieron públicamente para planificar su resistencia.”

EN: “The strike officially began on May 29, and on June 1 the manufacturers met publicly to plan their resistance.”

ES: “El 29 de mayo empezó la huelga. Los fabricantes se reunieron públicamente para planificar su respuesta el 1 de junio. Tenían dos estrategias.”

EN: “The strike officially began on May 29, and on June 1 the manufacturers met publicly to plan their resistance. Their strategies were carried out on two fronts.”

Cross-language plagiarism detection

ES: “La huelga comenzó oficialmente el 29 de mayo, y el 1 de junio los fabricantes se reunieron públicamente para planificar su resistencia.”

EN: “The strike officially began on May 29, and on June 1 the manufacturers met publicly to plan their resistance.”

ES: “El 29 de mayo empezó la huelga. Los fabricantes se reunieron públicamente para planificar su respuesta el 1 de junio. Tenían dos estrategias.”

EN: “The strike officially began on May 29, and on June 1 the manufacturers met publicly to plan their resistance. Their strategies were carried out on two fronts.”

Cross-Language Knowledge Graphs Analysis

CL-KGA [Franco-Salvador et al., 2013b] provides a context model by generating knowledge graphs from suspicious and source documents.

The similarity between two graphs G and G' is measured in a semantic graph space.

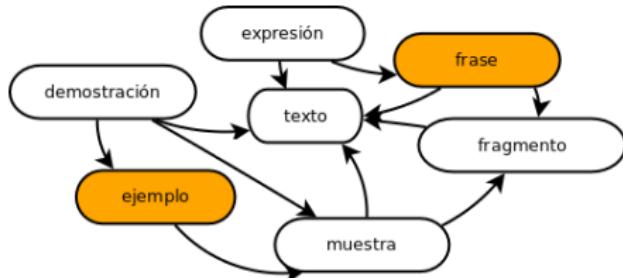
CL-KGA

The similarity between two graphs G and G' is measured in a semantic graph space.

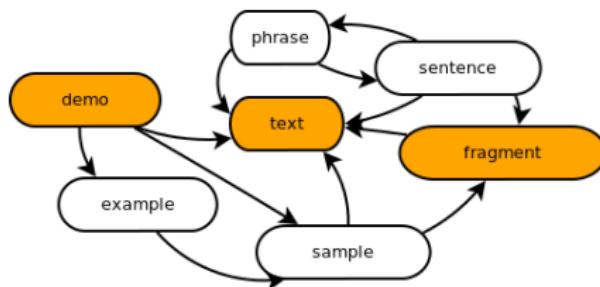
$$S(G, G') = S_c(G, G')(a + b S_r(G, G')) \quad (1)$$

$$S_c(G, G') = \frac{\sum_{c \in G \cap G'}^2 w(c)}{\sum_{c \in G} w(c) + \sum_{c \in G'} w(c)} \quad S_r(G, G') = \frac{\sum_{r \in N(c, G \cap G')}^2 w(r)}{\sum_{r \in N(c, G)} w(r) + \sum_{r \in N(c, G')} w(r)}$$

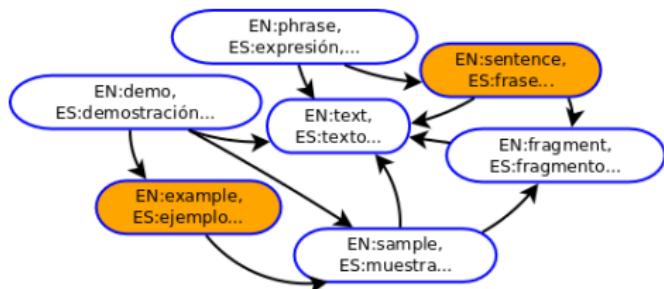
ES: “esta es una frase de ejemplo”



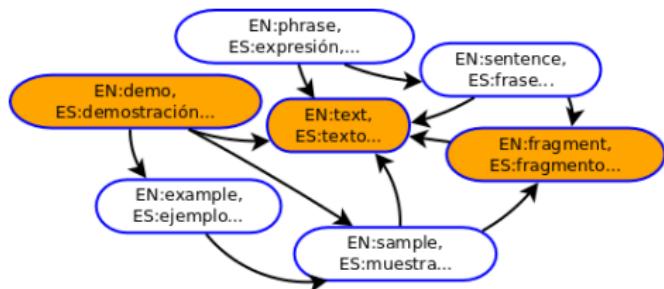
EN: “this is a **demo** text **fragment**”



ES: “esta es una frase de ejemplo”



EN: “this is a **demo** text fragment”

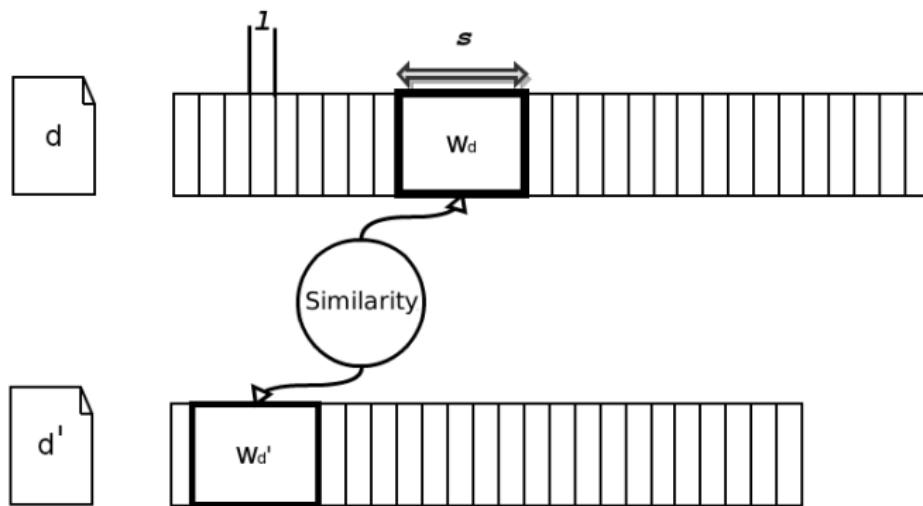


Cross-language Plagiarism Detection Evaluation

Task: Given a set of suspicious documents D and their corresponding source documents D' , the task is to compare pairs of documents (d, d') , $d \in D$ and $d' \in D'$, to find all plagiarized fragments in D from D' .

Cross-language Plagiarism Detection Evaluation

Task: Given a set of suspicious documents D and their corresponding source documents D' , the task is to compare pairs of documents (d, d') , $d \in D$ and $d' \in D'$, to find all plagiarized fragments in D from D' .



Cross-language Plagiarism Detection Evaluation

Dataset: We use the DE-EN and ES-EN cross-language plagiarism partition of PAN-PC'11 [Potthast et al., 2011b] competition.

Cross-language Plagiarism Detection Evaluation

Dataset: We use the DE-EN and ES-EN cross-language plagiarism partition of PAN-PC'11 [Potthast et al., 2011b] competition.

ES-EN documents	DE-EN documents
Suspicious 304	Suspicious 251
Source 202	Source 348
Plagiarism cases {ES,DE}-EN	
Automatic translation	5.142
Automatic translation + Manual correction	433

Cross-language Plagiarism Detection Evaluation

Models:

- CL-C3G
- CL-ASA_{IBMM1}
- CL-ASA_{BN}
- CL-KGA

Cross-language Plagiarism Detection Evaluation

DE-EN results:

Model	Plagdet	Recall	Precision	Granularity
CL-KGA	0.514	0.443	0.631	1.018
CL-ASA _{IBMM1}	0.406	0.344	0.604	1.113
CL-ASA _{BN}	0.289	0.222	0.595	1.172
CL-C3G	0.078	0.047	0.330	1.089

Cross-language Plagiarism Detection Evaluation

ES-EN results:

Model	Plagdet	Recall	Precision	Granularity
CL-KGA	0.599	0.525	0.703	1.004
CL-ASA _{BN}	0.554	0.491	0.663	1.030
CL-ASA _{IBMM1}	0.517	0.448	0.689	1.071
CL-C3G	0.170	0.128	0.617	1.372

Cross-language Plagiarism Detection Evaluation

Differences detecting automatic VS manual translations:

Cross-language Plagiarism Detection Evaluation

Differences detecting automatic VS manual translations:

Model	DE-EN				ES-EN			
	Recall		Precision		Recall		Precision	
	<i>automatic</i>	<i>manual</i>	<i>automatic</i>	<i>manual</i>	<i>automatic</i>	<i>manual</i>	<i>automatic</i>	<i>manual</i>
CL-KGA	0.538	0.247	0.698	0.098	0.601	0.221	0.774	0.098
CL-ASA _{IBMM1}	0.538	0.126	0.642	0.041	0.596	0.180	0.741	0.068
CL-ASA _{BN}	0.472	0.092	0.631	0.033	0.599	0.198	0.720	0.076

Cross-language Plagiarism Detection Evaluation

Differences detecting automatic VS manual translations:

Model	DE-EN				ES-EN			
	Recall		Precision		Recall		Precision	
	<i>automatic</i>	<i>manual</i>	<i>automatic</i>	<i>manual</i>	<i>automatic</i>	<i>manual</i>	<i>automatic</i>	<i>manual</i>
CL-KGA	0.538	0.247	0.698	0.098	0.601	0.221	0.774	0.098
CL-ASA _{IBMM1}	0.538	0.126	0.642	0.041	0.596	0.180	0.741	0.068
CL-ASA _{BN}	0.472	0.092	0.631	0.033	0.599	0.198	0.720	0.076

Number of manual cases ↓↓

Corpora Overview

corpus	focus	domain	language(s)	real	simulated	synthetic	annotated	CL
METER	journalistic re-use	press (politics & show-business)	en	■			□	
Co-derivatives	co-derivation in Wikipedia	multiple (encyclopedic)	de, en, es, hi	■				
PAN-PC	plagiarism	multiple	de, en, es	■	■	■	□	
CL!TR	CL re-use	CS and tourism (encyclopedic)	en, hi	■			■	

Corpora Overview

corpus	focus	domain	language(s)	real	simulated	synthetic	annotated	CL
METER	journalistic re-use	press (politics & show-business)	en	■			□	
Co-derivatives	co-derivation in Wikipedia	multiple (encyclopedic)	de, en, es, hi	■				
PAN-PC	plagiarism	multiple	de, en, es	■	■	■	□	
CL!TR	CL re-use	CS and tourism (encyclopedic)	en, hi	■			■	
JRC-Acquis	parallel	legislative	22 lan					
Wikipedia	comparable	encyclopedic	+250 lan					
Software Consumer	t. memory t. memory	technical magazine	en, eu + es, eu +					

■fully

□partially

Corpora Overview

corpus	focus	domain	language(s)	real	simulated	synthetic	annotated	CL
METER	journalistic re-use	press (politics & show-business)	en	■				□
Co-derivatives	co-derivation in Wikipedia	multiple (encyclopedic)	de, en, es, hi	■				
PAN-PC	plagiarism	multiple	de, en, es	■	■	■		□
CL!TR	CL re-use	CS and tourism (encyclopedic)	en, hi	■			■	
JRC-Acquis	parallel	legislative	22 lan					
Wikipedia	comparable	encyclopedic	+250 lan					
Software Consumer	t. memory t. memory	technical magazine	en, eu + es, eu +					

■fully

□partially

Outline

Introduction

Basic Concepts

Intrinsic Plagiarism Detection

External Plagiarism Detection

PAN Task on Plagiarism Detection

Plagiarism and Paraphrasing

Cross-Language Plagiarism Detection

PAN Tasks @ CLEF

Detection of Plagiarism in Source Code

PAN Competitions

PAN is a network around digital text forensics.

Mission

- Foster research and development in our tasks
- Push the limits of evaluating them
- Improve methodology for lab-style evaluations

Tasks

- Author Profiling (new in 2013)
- Author Identification
- Plagiarism Detection
 - Source Retrieval
 - Text Alignment

Software Submissions

- Instead of run submissions (i.e., software output on a given input)
- Improves sustainability, replicability, and reproducibility
- Increases participant engagement
- Allows for cross-year evaluations

Author Profiling

- Given a document, what are its author's demographics?

Corpus

- Genre: social media
- Languages: English, Spanish
- Size: 346 100 authors
- Annotations: age, gender

Selected results

- 21 softwares submitted
- Gender difficult to be discriminated, somewhat better in Spanish
- Age correctly detected in about 2/3 of cases

Award from the ForensicLab of the Universitat Pompeu Fabra

Author Identification

- Given a document, who wrote it?

Corpus

- Genres: non-fiction writing, short fiction, news
- Languages: English, Spanish, Greek
- Size: 120 cases
- Annotations: authorship

Selected results

- 18 softwares submitted
- Greek more difficult than English and Spanish
- Balancing performance in all languages with a single approach difficult
- Meta-model competitive to participants, but does not dominate

Plagiarism Detection

- Given a document, is it an original?

Corpus

- Genre: web, news
- Language: English
- Size: 10000 suspicious documents
- Annotations: reused text passages, obfuscation

Selected results

- 19 softwares submitted
- Advanced evaluation framework for web-scale retrieval
- Different retrieval paradigms open up trade-off between costs and recall
- Summary plagiarism most difficult to be detected
- First-time cross-year evaluation; first steps toward all-time evaluation

Software Submissions

Challenges Approaches

① Environment diversity virtualization

Support a wide variety of programming languages and operating systems.

② Executing untrusted software virtualization

Better be safe than sorry when executing binaries from a third party.

③ Data leakage sandboxing

Prevent data leaking by running software in a secured environment.

④ Error handling unit testing

Streamline the development round-trips for fixing execution errors.

⑤ Responsibility staged submissions

Incentivize participants to submit early.

⑥ Execution cost provide hardware or raise usage fees

We provided four servers each hosting up to 20 virtual machines.

Software Submissions

The 2013 Experience

- Entire lab accepts software submissions
- 62 virtual machines requested and provisioned
- 47 softwares installed, prepared for execution, and submitted by participants
- Testing and round-trips to fix errors
- Managed execution and evaluation using TIRA

The 2012 Experience

- One task accepts software submissions
- 10 softwares submitted
- Manual preparation for execution by us
- Testing and round-trips to fix errors
- Managed execution and evaluation using TIRA

<http://tira.webis.detira.webis.de>

Software Submissions

Error Analysis

- 1493 mails exchanged in 392 conversations
- 39 of 46 teams experienced at least one error, 26 at least two, 1 team 10
- No one panicked
- Staged submissions helped resolve errors early on
Rigorous unit testing and tools to assist participants in development

Software Submissions

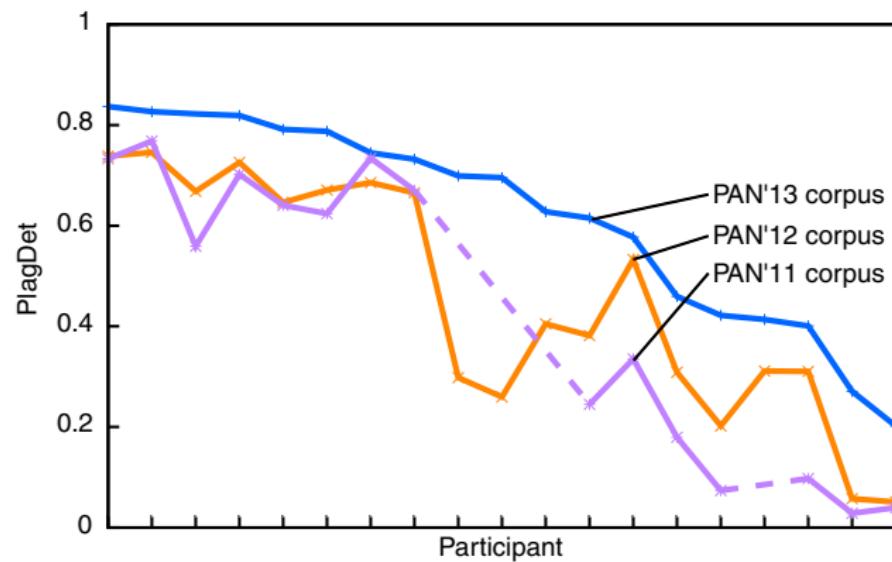
Cross-year Evaluation 2011-2013

Software Submission		PlagDet on PAN Plagiarism Corpus		
Team	Year	2013	2012	2011
Kong	2012	0.84	0.74	0.73
Oberreuter	2012	0.83	0.75	0.77
R. Torrejón	2013	0.82	0.67	0.56
Kong	2013	0.82	0.73	0.70
Palkovskii	2012	0.79	0.65	0.64
R. Torrejón	2012	0.79	0.67	0.62
Suchomel	2013	0.74	0.69	0.73
Suchomel	2012	0.73	0.67	0.67
Saremi	2013	0.70		
Shrestha	2013	0.70		
Kueppers	2012	0.63	0.40	
Palkovskii	2013	0.62	0.38	0.25
Nourian	2013	0.58	0.53	0.34
Sánchez-Vega	2012	0.46	0.31	0.18
Baseline		0.42	0.20	0.07
Gillam	2012	0.41	0.31	0.10
Gillam	2013	0.40	0.31	0.10
Jayapal	2013	0.27	0.06	0.03
Jayapal	2012	0.20	0.05	0.04

Software Submissions

Cross-year Evaluation 2011-2013

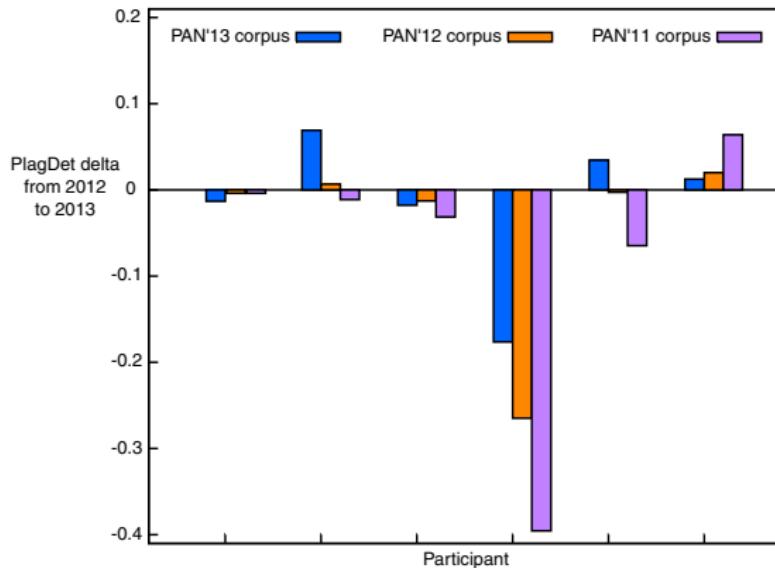
Assessing corpus difficulty



Software Submissions

Cross-year Evaluation 2011-2013

Assessing improvements across versions



Summary

Statistics	ALLC	SEPLN	FIRE			CLEF			
	2004	2009	2011	2012	2013	2010	2011	2012	2013
Task(s)	1	1	1	1	1	2	3	3	3
Follower		78				151	181	232	286
Registrations	11	21	6	12	16	53	52	68	110
Runs/Software	13	14	6	8	8	27	27	48	58
Notebooks	8	11	6	2	6	22	22	34	47
Attendees	5	18	6	30		25	36	61	

Take-away messages

- Software submissions improve sustainability
- Software submissions allow for re-evaluation
- Software submissions allow for cross-year evaluation
- Software submissions do not discourage participation

Outline

Introduction

Basic Concepts

Intrinsic Plagiarism Detection

External Plagiarism Detection

PAN Task on Plagiarism Detection

Plagiarism and Paraphrasing

Cross-Language Plagiarism Detection

PAN Tasks @ CLEF

Detection of Plagiarism in Source Code

Software Plagiarism

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

[Parker and Hamblen, 1989]

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

[Parker and Hamblen, 1989]

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]

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...

[Whale, 1986]

Software Plagiarism

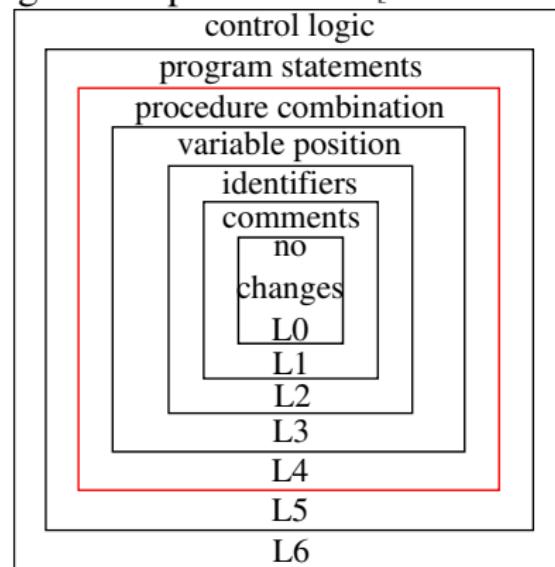
Techniques to disguise plagiarism

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]

Software Plagiarism

Program plagiarism spectrum [Faidhi and Robinson, 1987]



Plagiarism

Some (statistical) features

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		█

[Parker and Hamblen, 1989]

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]

Source Code Analysis Tools

MOSS ✓

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

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

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)

CL Source Code Analysis

Cross-language plagiarism makes sense in programming languages?

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...

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

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) {  
    $comment = "This is terrible";  
}  
elseif ($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) {  
    $comment = "This is terrible";  
}  
elseif ($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
```



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”

[Arwin and TahaGhoghi, 2006]

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]

Detection Process

- Intermediate code generation
- Filtering process (just a set of keywords is considered relevant)
- Comparison

Detection Process

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

Actual CL Analysis in Source Code?

Java

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

python

```
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";  
}
```



C++

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

perl

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

ASP

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

Ruby

```
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
```

#!/sh

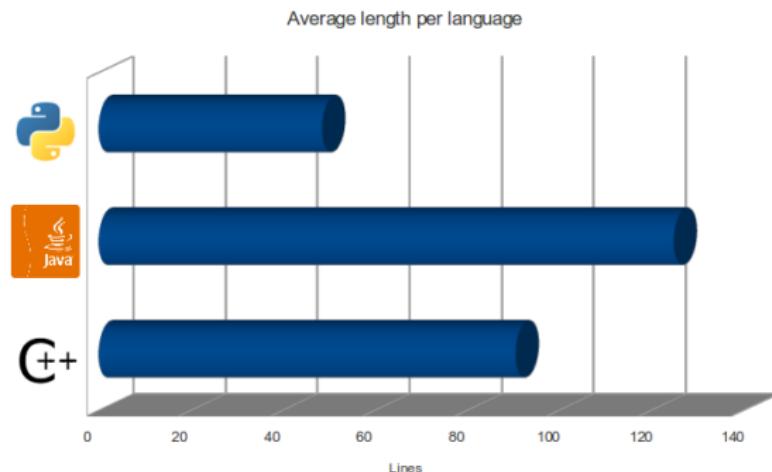


Actual CL Analysis in Source Code?

- ① Is there a length factor between programming languages?
 - C and Java lengths are closed...
 - Python is shorter...

Length factor

- Source codes extracted from the Qualification Round of Google Code Jam 2012 (<http://code.google.com/codejam/>)
- ~300 % of lines more between Java and Python
- ~40 % of lines more between Java and C++



[Flores et al., 2015]

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 Could we use a method such as CL-ESA?



ROSETTACODE.ORG

- Solutions to the same task in as many different programming languages as possible
- 745 tasks (Python-C share 546 tasks, Java-C 424, Java-Python 433)
- 546 programming languages

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 Could we use a method such as CL-ESA?
- ③ Is it possible to learn a bilingual dictionary of programming languages?
 - print printf 0.9; print echo 0.05...

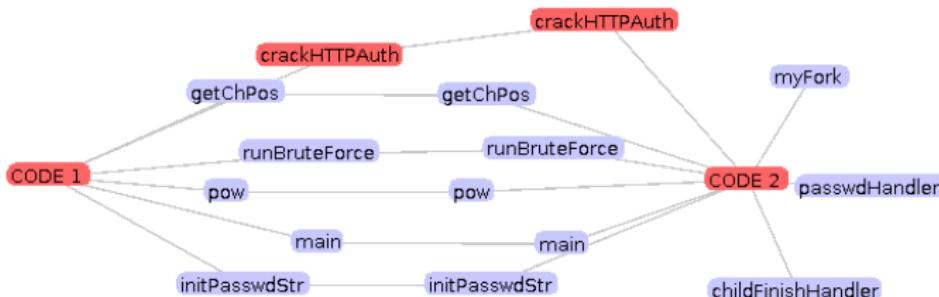


- Solutions to the same task in as many different programming languages as possible
- 745 tasks (Python-C share 546 tasks, Java-C 424, Java-Python 433)
- 546 programming languages
- **Manual labeling to distinguish between comparable and parallel source codes**
- **Creating artificial translations with source code translators**
- **Bilingual dictionary learned using IBM model 1**

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 Could we use a method such as CL-ESA?
- ③ Is it possible to learn a bilingual dictionary of programming languages?
 - print printf 0.9; print echo 0.05...
- ④ BTW: What about plagiarised methods/functions? (not entire programs)

DeSoCoRe



- Comparison at function/method level
- Manual similarity threshold
- CL comparison using character 3-grams

<http://memex2.dsic.upv.es/DeSoCoRe>

[Flores et al., 2012]

Task on Detection of SOurce COde Re-use

Detection of SOurce COde Re-use (SOCO)
FIRE 2014 Forum for Information Retrieval Evaluation
5 - 7 December 2014, Bangalore, India
<http://www.dsic.upv.es/grupos/nle/soco/>



Cross-Language detection of SOurce COde Re-use (CL-SOCO)
FIRE 2015 Forum for Information Retrieval Evaluation
4 - 6 December 2015, Gandhinagar, India
<http://www.dsic.upv.es/grupos/nle/clsoco/>

Plagiarism Detection

Thanks!

Paolo Rosso

prosso@dsic.upv.es

<http://www.dsic.upv.es/~prosso>

Natural Language Engineering Lab
PRHLT Research Center
Universitat Politècnica de València, Spain

Text Mining - Diploma on Big Data, ETSINF
20/06/15



Language Technologies
Natural Language Processing
NLEL
Natural Language Engineering Lab



References I



(2010).

Beijing, China.



Arwin, C. and TahaGhoghi, S. (2006).

Plagiarism Detection across Programming Languages.

In Proceedings of the Australasian Computer Science Conference (ACSC 2006), Tasmania, Australia.



Association of Teachers and Lecturers (2008).

School Work Plagued by Plagiarism - ATL Survey.

Technical report, Association of Teachers and Lecturers, London, UK.

Press release.



Barrón Cedeño, A. (2008).

Detección automática de plagio en texto.

Master's thesis, Universidad Politécnica de Valencia, Valencia, España.

Advisor: Paolo Rosso.



Barrón-Cedeño, A. (2012).

On the mono- and cross-language detection of text re-use and plagiarism.

PhD thesis, Universitat Politècnica de València, Valencia, Spain.



Barrón-Cedeño, A., Gupta, P., and Rosso, P. (2013a).

Methods for cross-language plagiarism detection.

Knowledge-Based Systems, 50:211–217.



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. Springer-Verlag.

References II



Barrón-Cedeño, A., Rosso, P., Agirre, E., and Labaka, G. (2010).

Plagiarism Detection across Distant Language Pairs.

In [col, 2010].



Barrón-Cedeño, A., Rosso, P., and Benedí, J. (2009).

Reducing the Plagiarism Detection Search Space on the Basis of the Kullback-Leibler Distance.

Computational Linguistics and Intelligent Text Processing, 10th International Conference, LNCS (5449):523–534. Springer-Verlag.



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, Patras, Greece. CEUR-WS.org.



Barrón-Cedeño, A., Rosso, P., Pinto, D., and Juan, A. (2008).

On Cross-lingual Plagiarism Analysis using a Statistical Model.

In Proceedings of the ECAI'08 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse, PAN'08.



Barrón-Cedeño, A., Vila, M., Martí, M. A., and Rosso, P. (2013b).

Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection.

Computational Linguistics, 39(4):917–947.



Barzilay, R. and Lee, L. (2003).

Learning to Paraphrase: An Unsupervised Approach Using Multiple-Sequence Alignment.

In Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL 2003), Edmonton, Canada. ACL.

References III



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.



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, Berkeley, CA. ACM.



Bernstein, Y. and Zobel, J. (2004).

A Scalable System for Identifying Co-Derivative Documents.

String Processing and Information Retrieval. Proceedings of the Symposium on String Processing and Information Retrieval, LNCS (3246):1–11.

Springer-Verlag.



Bigi, B. (2003).

Using Kullback-Leibler Distance for Text Categorization.

Advances in Information Retrieval: Proceedings of the 25th European Conference on IR Research (ECIR 2003), LNCS (2633):305–319.

Springer-Verlag.



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.

References IV



Broder, A. (1997).

On the Resemblance and Containment of Documents.

In Compression and Complexity of Sequences (SEQUENCES'97), pages 21–29, Salerno, Italy. 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.



Burrows, S., Potthast, M., and Stein, B. (2012).

Paraphrase Acquisition via Crowdsourcing and Machine Learning (to appear).

ACM Transactions on Intelligent Systems and Technology.



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.



Chapman, K. and Lupton, R. (2004).

Academic Dishonesty in a Global Educational Market: A Comparison of Hong Kong and American University Business Students.

International Journal of Educational Management, 18(7):425–435.



Clough, P. (2003).

Old and new challenges in automatic plagiarism detection.

National UK Plagiarism Advisory Service.



Clough, P. and Gaizauskas, R. (2009).

Corpora and Text Re-Use.

In Lüdeling, A., Kyö, M., and McEnery, T., editors, Handbook of Corpus Linguistics, Handbooks of Linguistics and Communication Science, pages 1249—1271. Mouton de Gruyter.

References V



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, Las Palmas, Spain.



Comas, R. and Sureda, J. (2008).

Academic Cyberplagiarism: Tracing the Causes to Reach Solutions.

In Comas, R. and Sureda, J., editors, Academic Cyberplagiarism [online dossier], volume 10 of Digithum. Iss, pages 1–6. UOC. [http://bit.ly/cyberplagiarism_cs].



Dolan, W. and Brockett, C. (2005).

Automatically Constructing a Corpus of Sentential Paraphrases.

In Proceedings of the Third International Workshop on Paraphrasing (IWP 2005), Jeju, Korea.



Faidhi, J. and Robinson, S. (1987).

An empirical approach for detecting program similarity and plagiarism within a university programming environment.

Comput. Educ., 11(1).



Flores, E., Barrón-Cedeño, A., Moreno, L., and Rosso, P. (2015).

Uncovering source code re-use in large-scale programming environments.

In Computer Applications in Engineering and Education, volume 23, pages 383–390.



Flores, E., Barrón-Cedeño, A., Rosso, P., and Moreno, L. (2012).

Descore: Detecting source code re-use across programming languages.

In Proceedings of the Demonstration Session at the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1–4, Montréal, Canada. Association for Computational Linguistics.

References VI



Franco-Salvador, M., Gupta, P., and Rosso, P. (2013a).

Cross-language plagiarism detection using a multilingual semantic network.

In Advances in Information Retrieval, pages 710–713. Springer.



Franco-Salvador, M., Gupta, P., and Rosso, P. (2013b).

Cross-language plagiarism detection using a multilingual semantic network.

In Proc. of the 35th European Conference on Information Retrieval (ECIR'13), volume LNCS(7814). Springer-Verlag.



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.



Grman, J. and Ravas, R. (2011).

Improved Implementation for Finding Text Similarities in Large Collections of Data - Notebook for PAN at CLEF 2011.

In [Petras et al., 2011].



Grozea, C., Gehl, C., and Popescu, M. (2009).

ENCOPILOT: Pairwise Sequence Matching in Linear Time Applied to Plagiarism Detection.

In [Stein et al., 2009], pages 10–18.



Gupta, P., Barrón-Cedeno, A., and Rosso, P. (2012).

Cross-language high similarity search using a conceptual thesaurus.

In Information Access Evaluation. Multilinguality, Multimodality, and Visual Analytics, pages 67–75. Springer.



Haines, V., Diekhoff, G., LaBeff, G., and Clarck, R. (1986).

College Cheating: Inmaturity, Lack of Commitment, and the Neutralizing Attitude.

Research in Higher Education, 25(4):342–354.

References VII

-  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://bit.ly/ieee_plagiarism].
Published: 2008; Accessed 3/Mar/2010.
-  Irribarne, R. and Retondo, H. (1981).
Plagio de obras literarias. Ilícitos 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 de 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.
Text, Speech and Dialogue (TSD 2006), LNAI (4188):661–667.
Springer-Verlag.
-  Kasprzak, J. and Brandejs, M. (2010).
Improving the Reliability of the Plagiarism Detection. System Lab Report for PAN at CLEF 2010.
In [Braschler and Harman, 2010].

References VIII



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, Prague, Czech Republic.



Kulathuramaiyer, N. and Maurer, H. (2007).

Coping With the Copy-Paste-Syndrome.

In Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2007 (E-Learn 2007), pages 1072—1079, Quebec City, Canada. AACE.



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.



MacQueen, J. (1967).

Some Methods for Classification and Analysis of MultiVariate Observations.

In Cam, L. and Neyman, J., editors, Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, volume 1, pages 281–297. University of California Press.

References IX



Martin, B. (1994).
Plagairism: a Misplaced Emphasis.
Journal of Information Ethics, 3(2):36–47.



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



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., 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.
Springer-Verlag.



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].

References X



Navigli, R. and Ponzetto, S. P. (2012).
BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network.
Artificial Intelligence, 193:217–250.



Nawab, R., Stevenson, M., and Clough, P. (2010).
University of Sheffield Lab Report for PAN at CLEF 2010.
In [Braschler and Harman, 2010].



Oberreuter, G., L'Huillier, G., Rfos, S., and Velásquez, J. (2011).
Approaches for Intrinsic and External Plagiarism Detection: Notebook for PAN at CLEF 2011.
In [Petras et al., 2011].



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>].



Palkovskii, Y., Belov, A., and Muzika, I. (2011).
Using WordNet-based Semantic Similarity Measurement in External Plagiarism Detection - Notebook for PAN at CLEF 2011.
In [Petras et al., 2011].



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



Petras, V., Forner, P., and Clough, P., editors (2011).
Notebook Papers of CLEF 2011 LABs and Workshops, Amsterdam, The Netherlands.

References XI



Pinto, D., Civera, J., Barrón-Cedeño, A., Juan, A., and Rosso, P. (2009a).

A Statistical Approach to Crosslingual Natural Language Tasks.

Journal of Algorithms, 64(1):51–60.



Pinto, D., Civera, J., Barrón-Cedeño, A., Juan, A., and Rosso, P. (2009b).

A Statistical Approach to Crosslingual Natural Language Tasks.

J. Algorithms, 64(1):51–60.



Potthast, M., Barrón-Cedeño, A., Stein, B., and Rosso, P. (2011a).

Cross-Language Plagiarism Detection.

Language Resources and Evaluation (LRE), Special Issue on Plagiarism and Authorship Analysis, 45(1):1–18.



Potthast, M., Eiselt, A., Barrón-Cedeño, A., Stein, B., and Rosso, P. (2011b).

Overview of the 3rd int. competition on plagiarism detection.

In CLEF (Notebook Papers/Labs/Workshop).



Potthast, M., Stein, B., and Anderka, M. (2008).

A Wikipedia-Based Multilingual Retrieval Model.

Advances in Information Retrieval, 30th European Conference on IR Research, LNCS (4956):522–530.

Springer-Verlag.



Potthast, M., Stein, B., Barrón-Cedeño, A., and Rosso, P. (2010).

An Evaluation Framework for Plagiarism Detection.

In [col, 2010], pages 997–1005.



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.

References XII



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, Borovets, Bulgaria.



Pouliquen, B., Steinberger, R., and Ignat, C. (2006).

Automatic Annotation of Multilingual Text Collections with a Conceptual Thesaurus.

CoRR, abs/cs/0609059.



R. Costa-jussà, M., Banchs, R., Grivolla, J., and Codina, J. (2010).

Plagiarism Detection Using Information Retrieval and Similarity Measures based on Image Processing Techniques.

In [Braschler and Harman, 2010].



Recasens, M. and Vila, M. (2010).

On Paraphrase and Coreference.

Computational Linguistics, 36(4):639–647.



Rodríguez Torrejón, D. and Martín Ramos, J. (2010a).

CoReMo System (Contextual Reference Monotony).

In [Braschler and Harman, 2010].



Rodríguez Torrejón, 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.

References XIII



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, Graz, Austria. Know-Center.



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), San Sebastian, Spain. CEUS-WS.org.



Stolcke, A. (2002).

SRILM - An Extensible Language Modeling toolkit.

In Intl. Conference on Spoken Language Processing, Denver, Colorado.

References XIV

-  Taylor, F. (1965).
Cryptomnesia and Plagiarism.
The British Journal of Psychiatry, 111:1111–1118.
-  Vila, M., Martí, M., and Rodríguez, H. (2011).
Paraphrase Concept and Typology. A Linguistically Based and Computationally Oriented Approach.
Procesamiento del Lenguaje Natural, 46:83–90.
-  Weber, S. (2007).
Das Google-Copy-Paste-Syndrom. Wie Netzplagiäte Ausbildung und Wissen gefährden.
Telepolis.
-  Whale, G. (1986).
Detection of plagiarism in student programs.
In Proceedings of the Ninth Australasian Computer Science Conference (ACSC 1986), pages 231–241.
-  Wikipedia (2010a).
Hash.
[http://bit.ly/wikipedia_hash].
Accessed 17/Sep/2010.
-  Wikipedia (2010b).
Party of European Socialists | Partido Socialista Europeo | Europako Alderdi Sozialista.
[http://bit.ly/wikipedia_socialists].
Accessed 10/Feb/2010.