Detección automática de plagio en texto

L. Alberto Barrón Cedeño

Máster en inteligencia artificial, reconocimiento de formas e imagen digital

Advisor: Paolo Rosso

Natural Language Engineering Lab, ELiRF

Universidad Politécnica de Valencia

IV Jornadas MAVIR
November 19th, 2009

Outline

What is Plagiarism?

- **Copying words or ideas** from someone else without giving credit.
- **Changing words** but copying the sentence structure of a source without giving credit.
- Copying so many words or ideas from a source that it makes up the majority of your work, whether you give credit or not.

www.plagiarism.org

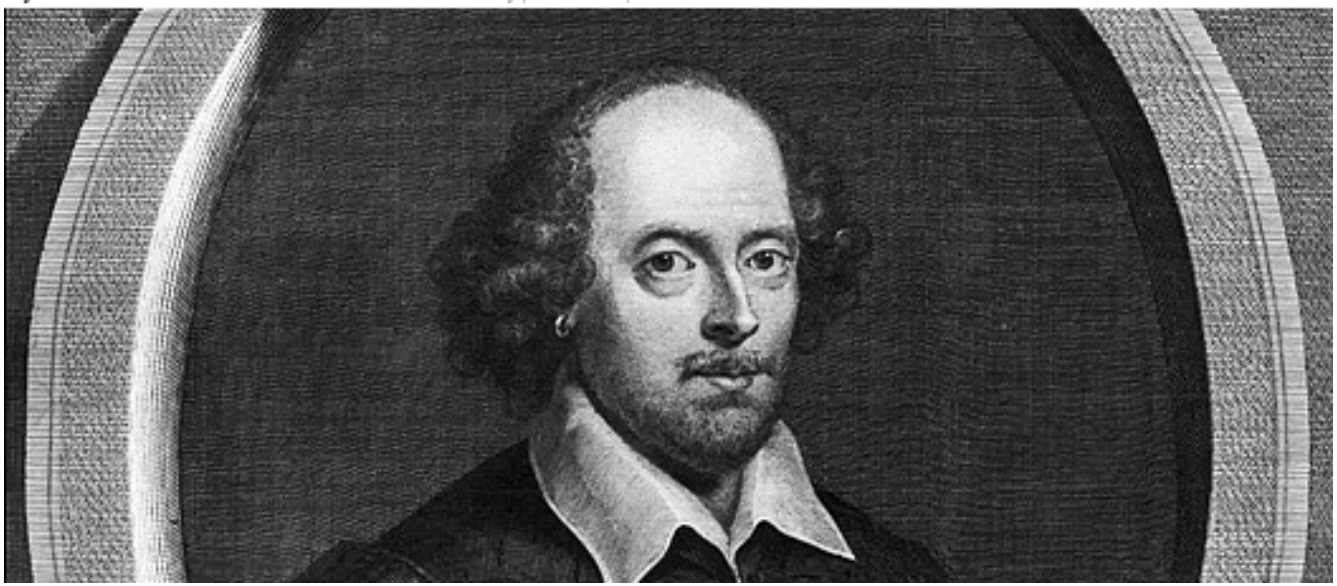
Detention vs. Detection

“En un lugar de la **estancia**, de cuyo nombre no quiero acordarme...”

In the news

Plagiarism Software Finds a New Shakespeare Play

By **GAELLE FAURE / LONDON** Tuesday, Oct. 20, 2009



Some available (online) systems:

- PI@giarism
- turnitin
- wCopyFind
- doccop.com
- Ferret
- Plagiarismdetect

http://www.time.com (Oct. 20, 2009)

Objectives

- Study of some of the current (statistical) approaches to plagiarism detection.
- Analysis and adaptation of available lexical and statistical resources
- Becoming the seed of a broader (PhD) research on **monolingual** and **cross-lingual** plagiarism detection.

Terminology

$\mathcal{A}$	author
d_q	plagiarism suspicious document
d	potential source document
D / D_q	set of source/suspicious documents
s	text fragment in a document

Relevant Factors for Plagiarism Detection

intrinsic analysis	Use of vocabulary Changes of vocabulary Punctuation Readability of text
external analysis	Amount of similarity between texts Distribution of words

[Clough, 2000]

“Search and classify”

d_q	determine whether d_q was written by one single. If not, identify sections written by a different author.
d_q to d	retrieve documents $d \in D$ such that d may be the source of the potentially plagiarised d_q
$s \in d_q$ to d	For (some) section $s \in d_q$, retrieve documents $d \in D$ such that d may be the source of potentially plagiarised sections in d_q
$s \in d_q$ to $s \in d$	For (some) section $s \in d_q$, retrieve sections $s \in d \in D$ such that $s \in d$ may be the source of the potentially plagiarised section $s \in d_q$

Outline

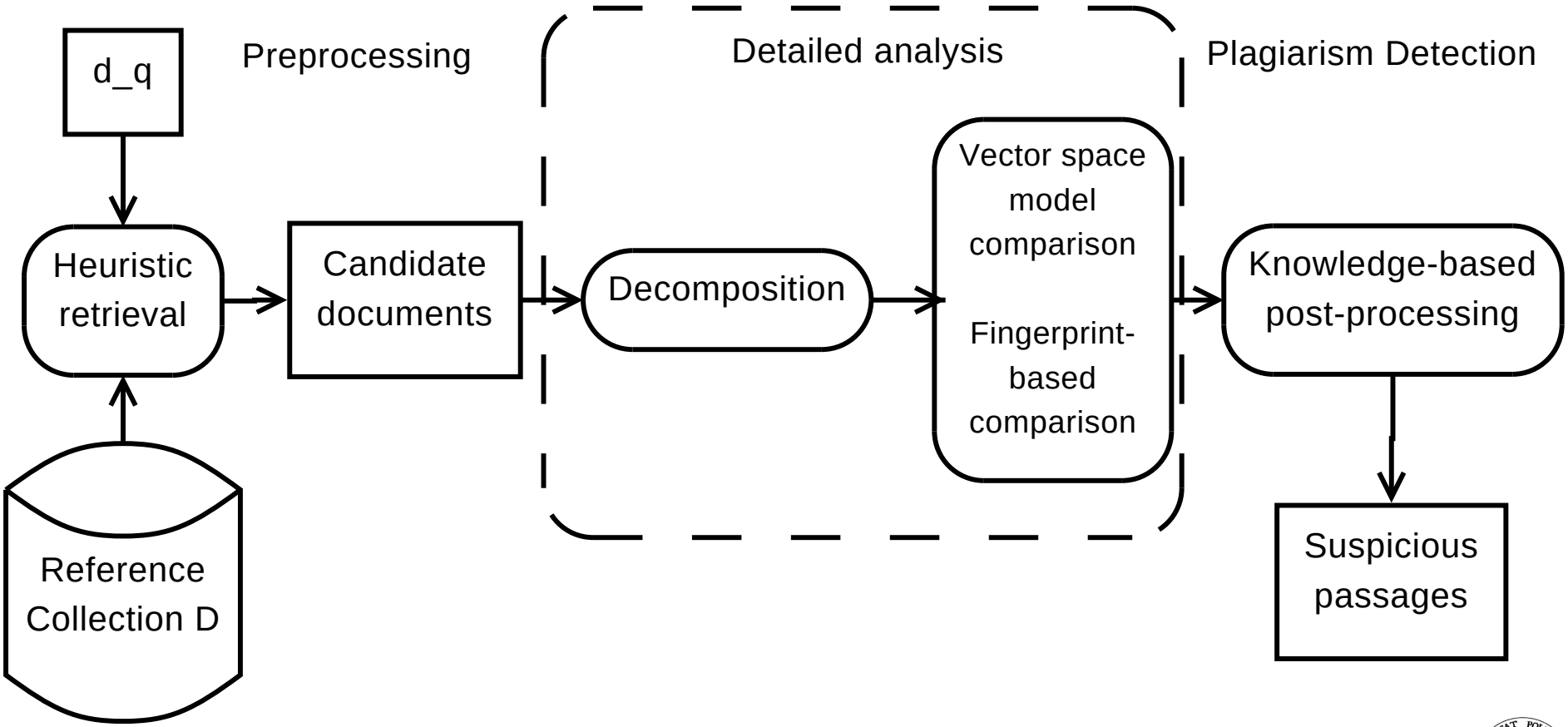
Intrinsic Analysis

If a person is able to detect a plagiarism case by reading a text…

- Word average frequency class
- Average [sentence , word] length
- Stopwords average
- Complexity measures

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

Overview of External Analysis



(adapted from [Potthast et al., 2009a, LRE, submitted])

External Analysis

n-gram based comparison

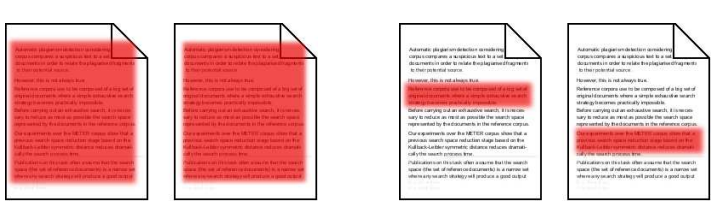
Given $N(x)$, the set of n-grams in x

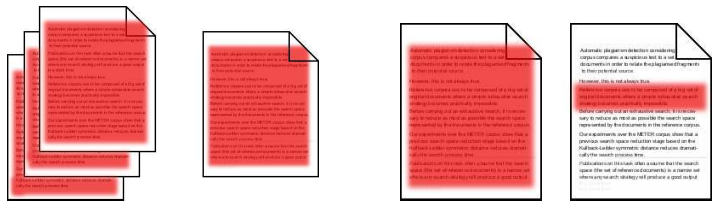
Resemblance

$$R(d_q \mid d) = \frac{|N(d_q) \cap N(d)|}{|N(d_q) \cup N(d)|}$$

Containment

$$C(s \in d_q \mid d) = \frac{|N(s) \cap N(d)|}{|N(s)|}$$





[Broder, 1997, Lyon et al., 2001]

Why do n-grams work? (1)

Query	|web pages|
Bienvenido	57,800,000
Bienvenido a	30,000,000
Bienvenido a la	10,700,000
Bienvenido a la página web	572,000
Bienvenido a la página web del	265,000
Bienvenido a la página web del posgrado	5
Bienvenido a la página web del posgrado en informática, desde	4

Sentence splitting does not work (cf. www.plagiarismdetect.com/)

http://www.popinformatica.upv.es/ (1-XII-08)

(Example adapted from P. Clough)

Why do n-grams work? (2)

- Consider 4 documents $d_{1...4}$
- $d_{1...4}$ were authored by $\mathcal{A}$
- $d_{1...4}$ have a common topic
- Avg. document size: $3,728$ words

$\lvert d \rvert$	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

Fingerprinting models

- D is compiled into a fingerprint index
- fingerprints of d_q and d are compared

- COPS [Brin et al., 1995]
- Winnowing [Schleimer et al., 2003]
- Fuzzy fingerprinting [Stein, 2007]

Outline

Language Models

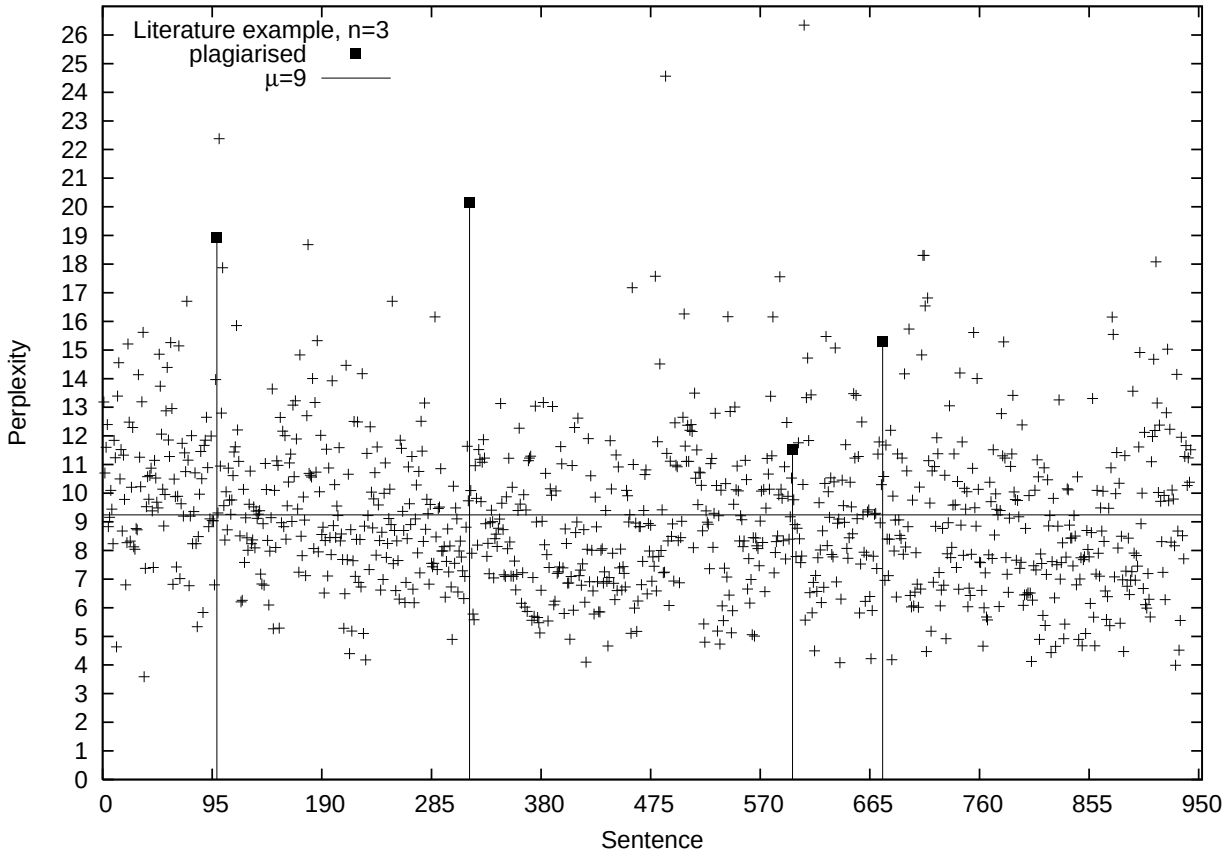
$$P(w_1, \ldots, w_k) = \prod_{i=1}^{k} P(w_i \mid w_1, \ldots, w_i - 1)$$

- Is it possible to characterize the writing style with a *LM*?
- Perplexity (PP) determines how well a LM predicts a text
- Compute a language model LM from $D_{\mathcal{A}}$.
- Let d_1 and d_2 be two documents such that $d_1 \in \mathcal{A}$ and $d_2 \notin \mathcal{A}$
- We **expect** that $PP(d_1) \ll PP(d_2)$.

[Barrón-Cedeño and Rosso, 2008, PAN-lm]

Language Models

POS



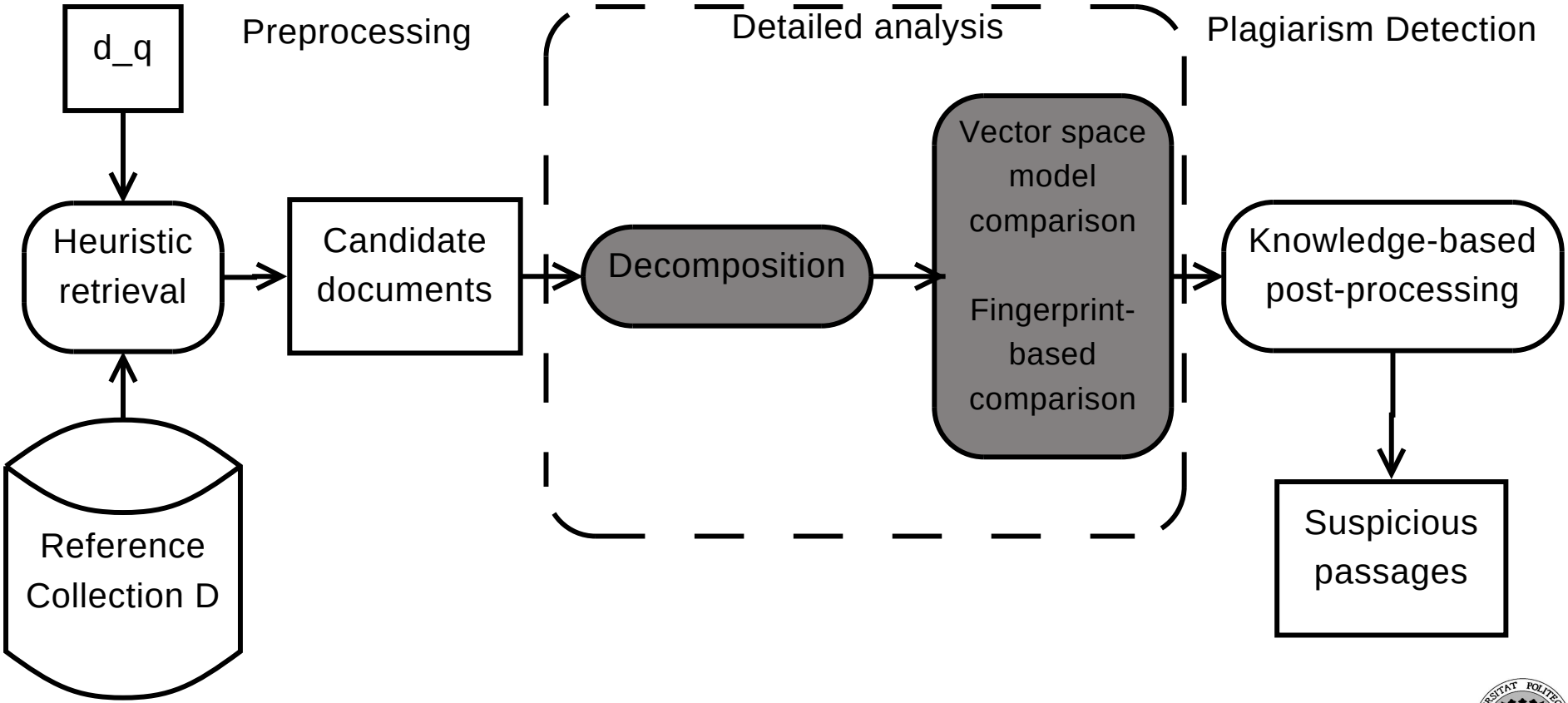
Language Models

What is wrong?

- In fact, this approach is closer to intrinsic analysis
- Significant text fragments contain more than 100 words
- Vectors (or LM) should be computed at character level

[Stamatatos, 2009]

Detailed Analysis



n-grams Comparison

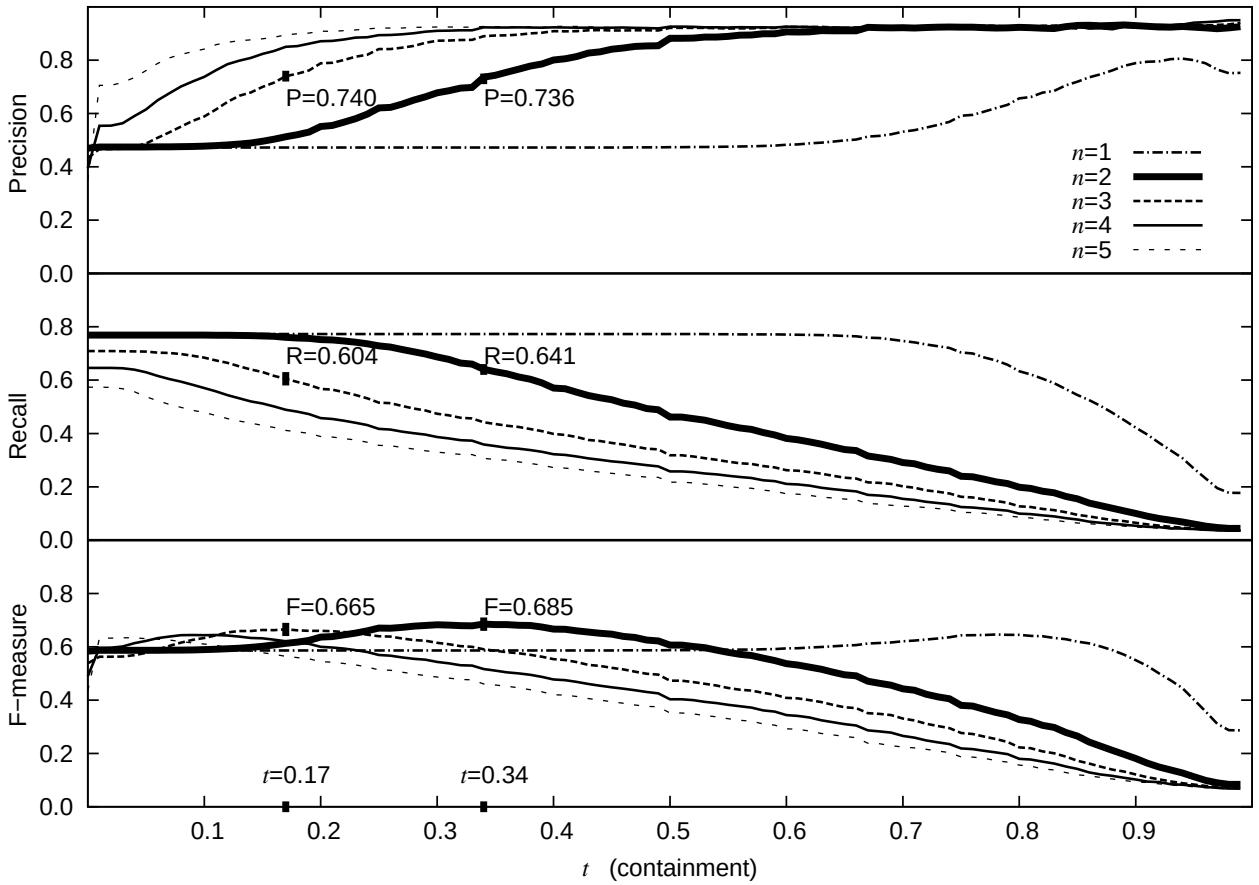
- Meter corpus on journalistic text reuse [Clough et al., 2002]
- Containment approach ($s \in d_q$ to d)

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

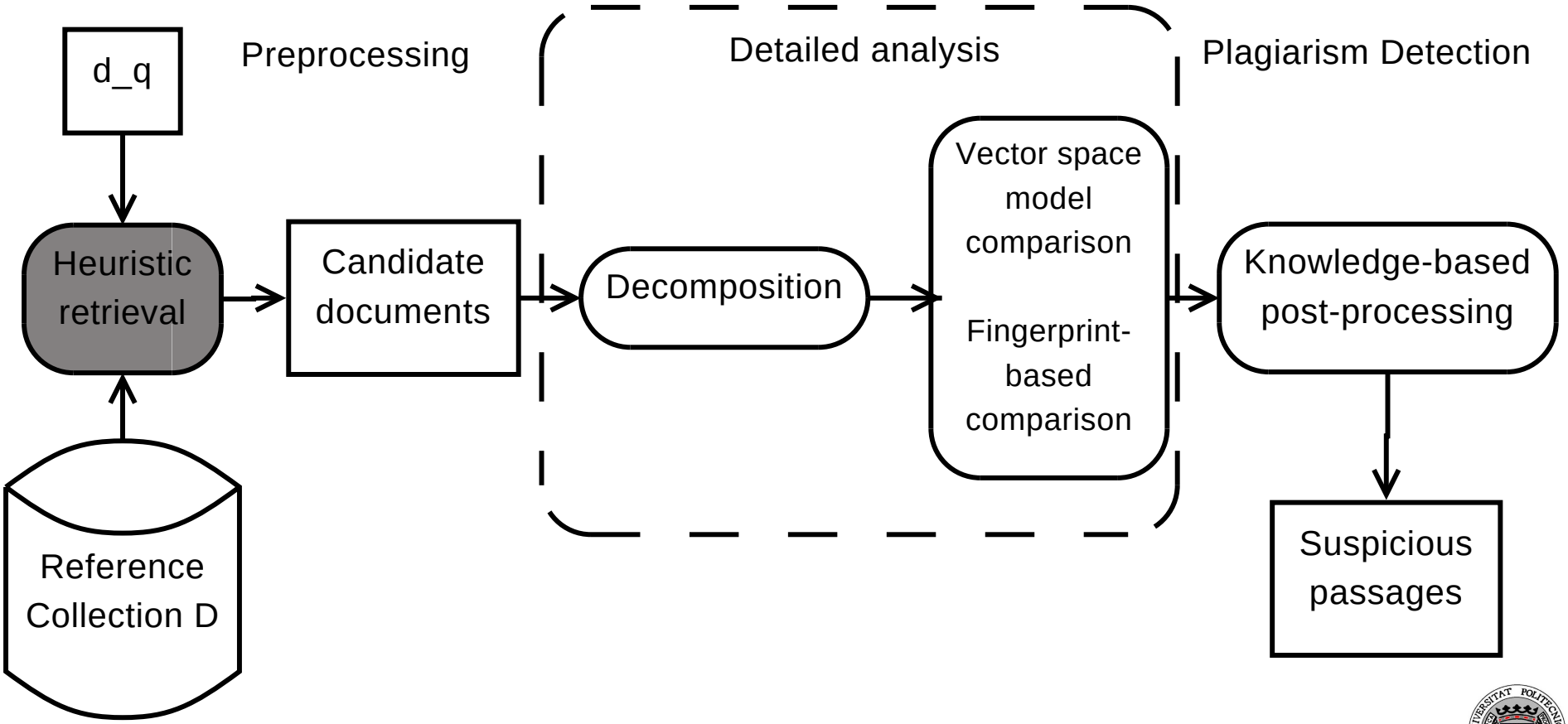
[Barrón-Cedeño and Rosso, 2009a, ECIR]

n-grams Comparison

How long n should be?

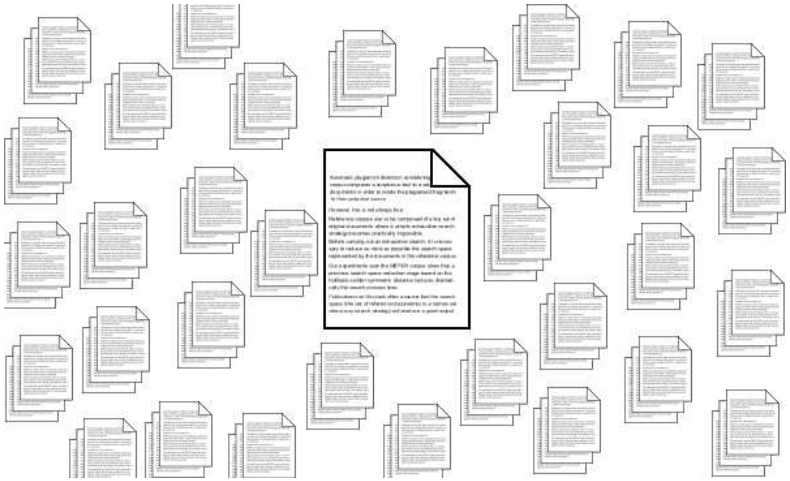


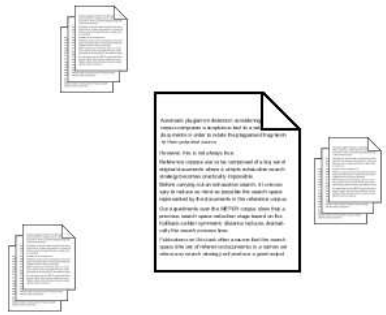
“Heuristic Retrieval”



Previous selection of (a good) D

- Some methods work properly. However...
  what about the size of D? (database, digital library, Internet)





Previous selection of (a good) D

Based on the Kullback-Leibler distance

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

[Kullback and Leibler, 1951, Bigi, 2003]

- Variation of n-gram levels $n = \{1, 2, 3\}$
- Keywords ranking on the basis of tf, $tfidf$ and tp
- The top 20% of the keywords ranking represents the document
- $P_d(t_i) = tf_{d,t_i}$; $Q_{d_q}(t_i) = (tf_{d_q,t_i} \mid P_d(t_i))$

[Barrón-Cedeño et al., 2009b, CICLing]

Previous selection of (a good) D

Results

- Best option: $tfidf$ on 1-grams
- Exhaustive comparison: Containmet on 3-grams

Selection of D	**threshold**	**P**	**R**	**F**	**t**
NO	0.34	0.73	0.63	0.68	2.32
YES	**0.25**	**0.77**	**0.74**	**0.75**	**0.19**

[Barrón-Cedeño and Rosso, 2009b, SEPLN]

Outline

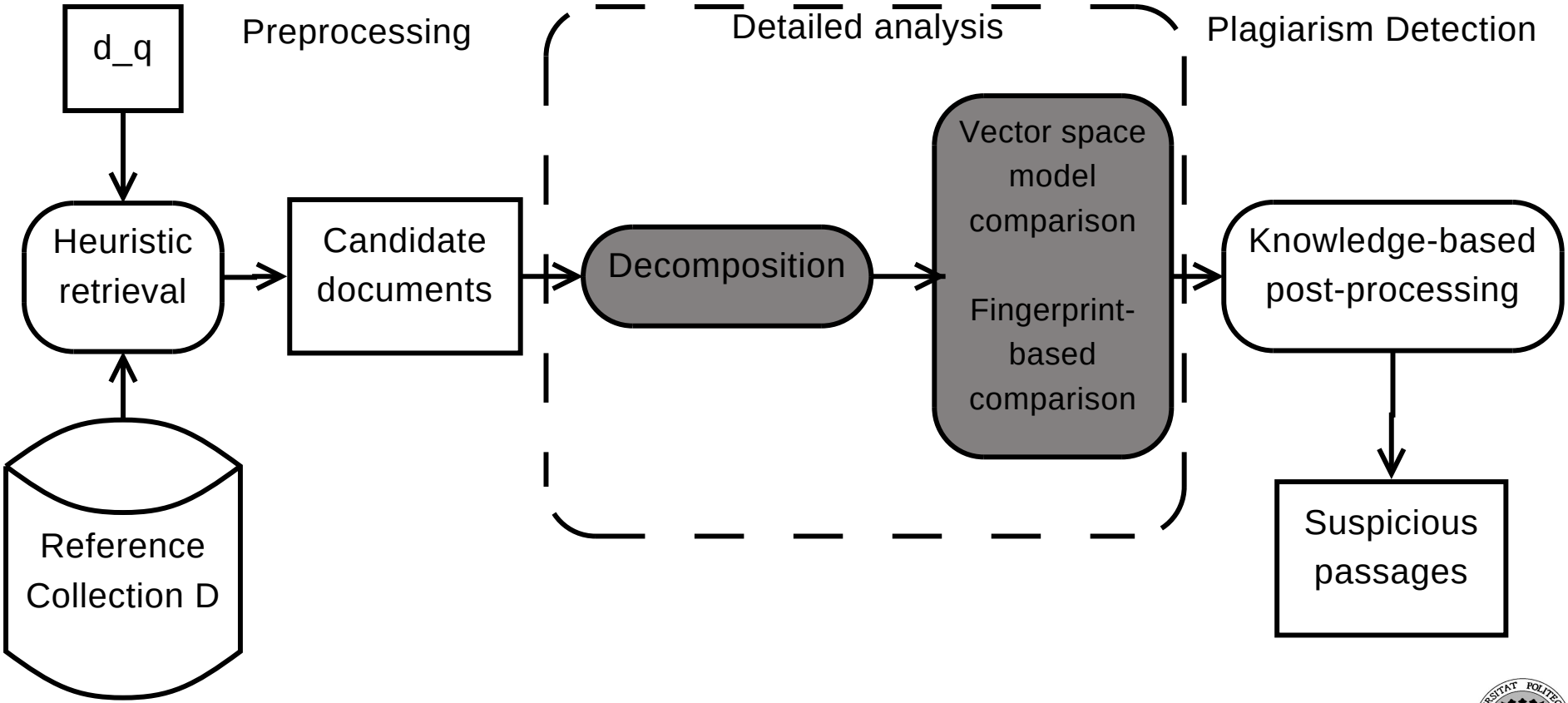
Cross-Language plagiarism detection

What to do if d_q and d are written in different languages?

- Scholars from non-English speaking countries write texts in their native languages
- Current scientific discourse to refer to is often published in English

Issue The syntactical similarity between passages in d_q and d is lost across languages.

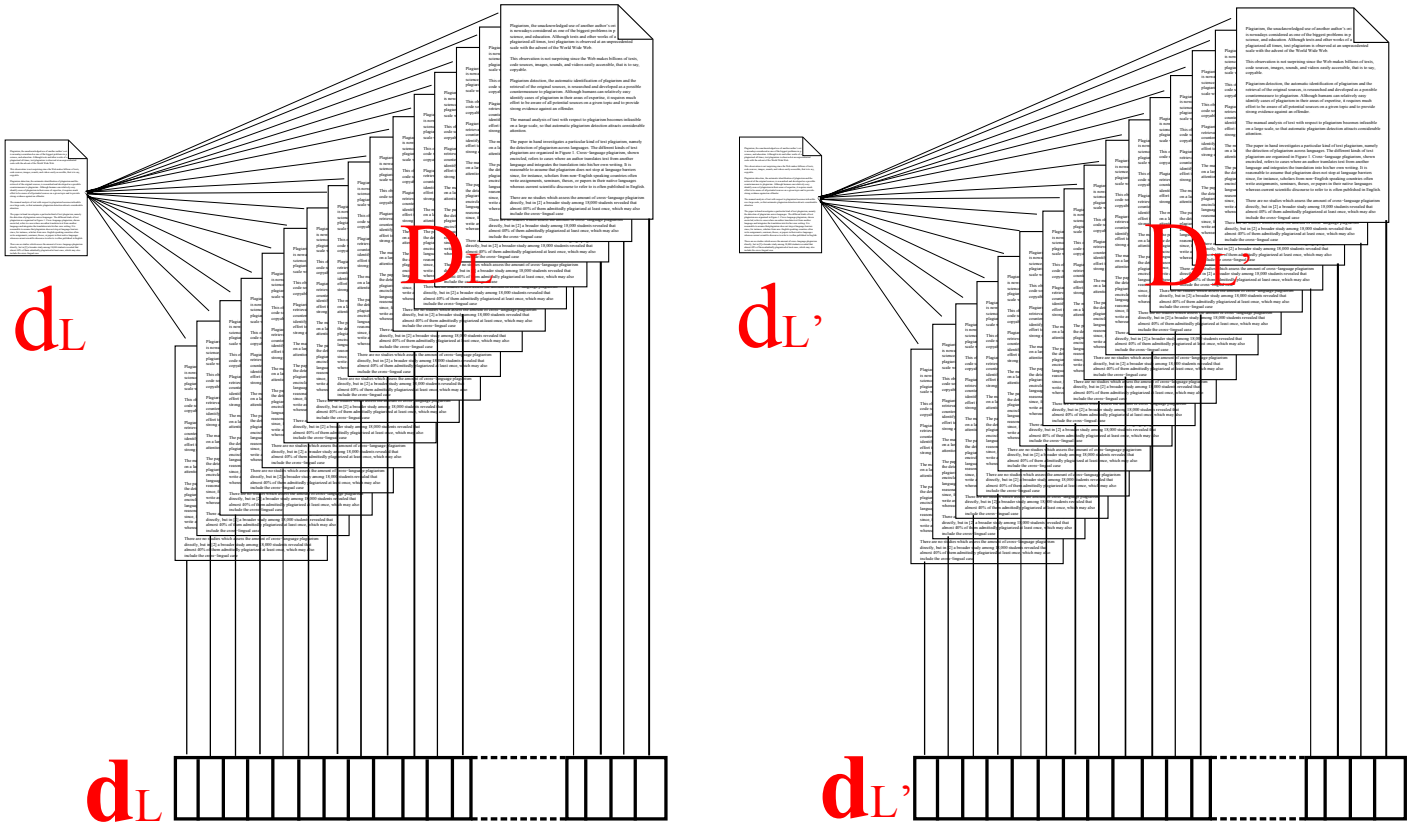
Cross-Language Plagiarism Detection



Cross-Language Plagiarism Detection

Some options:

1. EUROVOC Thesaurus-based [Pouliquen et al., 2003]
2. CL-ESA, Wikipedia-based [Potthast et al., 2008]



Cross-Language Plagiarism Detection

CL-ASA: CL-Alignment-based Similarity Analysis

- How probable is that d_q is a valid translation of d'?
- Combination of a two-step probabilistic translation and similarity analysis
- Adaptation of the basic principles of statistical Machine Translation's IBM M1 [Brown et al., 1990]

[Barrón-Cedeño et al., 2008, PAN-cl], [Pinto et al., 2009, Algorithms]

Cross-Language Plagiarism Detection

Bayes's rule for statistical Machine Translation [Brown et al., 1993]

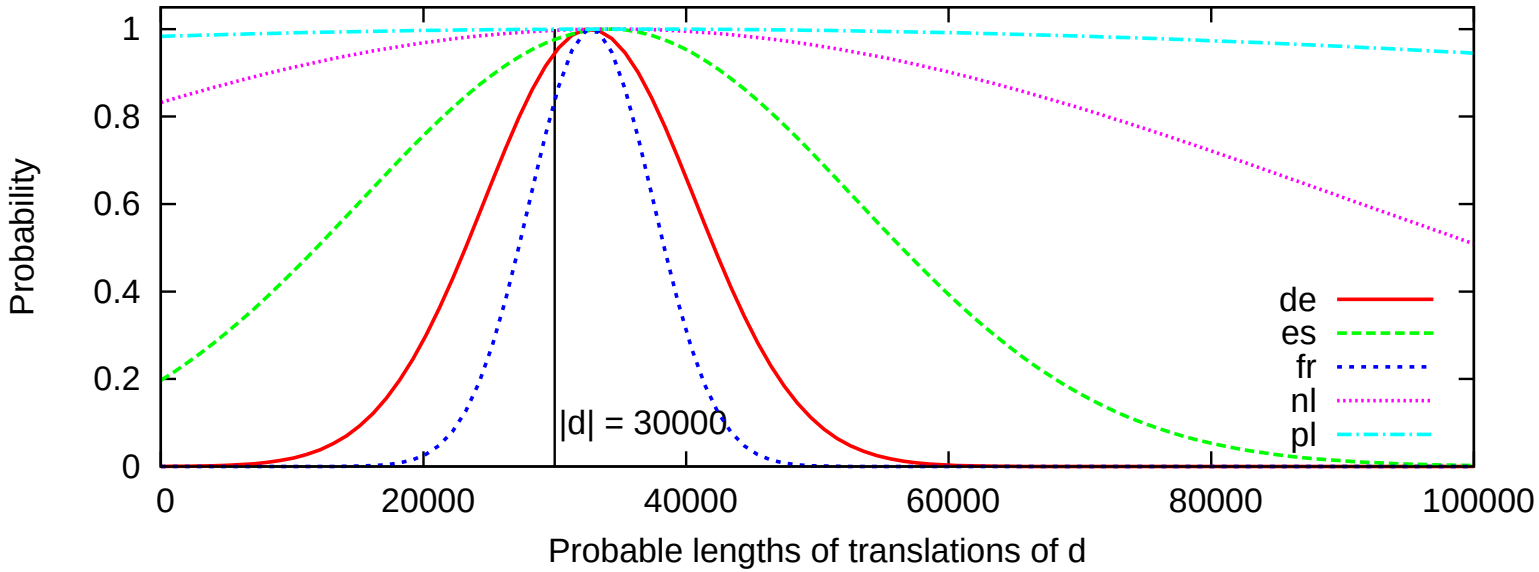
$$p(d' \mid d_q) = \frac{p(d')\ p(d_q \mid d')}{p(d_q)}$$

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

es	**en**	$p(es, en)$	es	**en**	$p(es, en)$
certifica	certifies	0.4203	certifica	certifying	0.0913
certifica	certify	0.1644	certifica	hereby	0.0548
certifica	certified	0.1096	...		

Cross-Language Plagiarism Detection

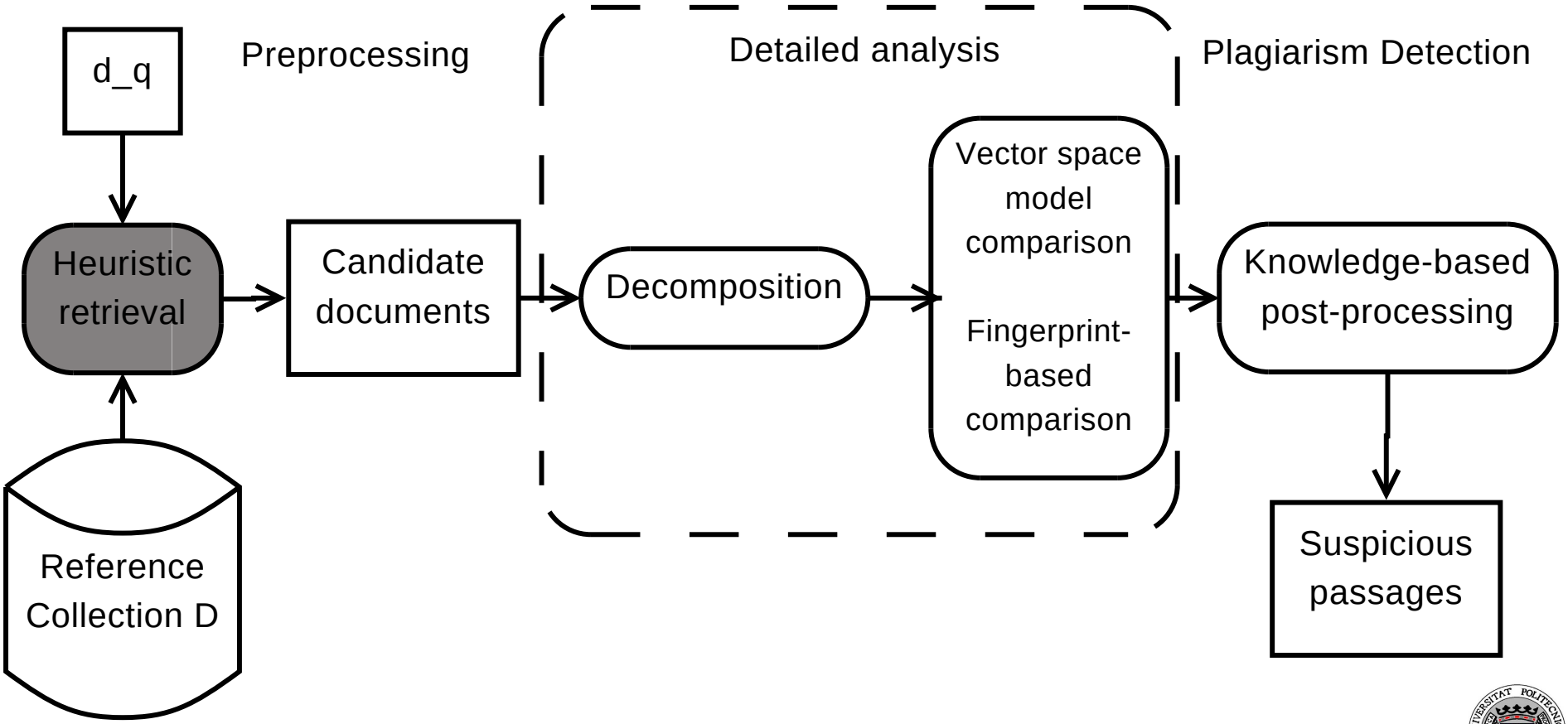
- . $p(d')$ is the *language model probability*
- Length model, inspired on [Pouliquen et al., 2003]



The tests in a toy corpus were really promising

Outline

Contributions



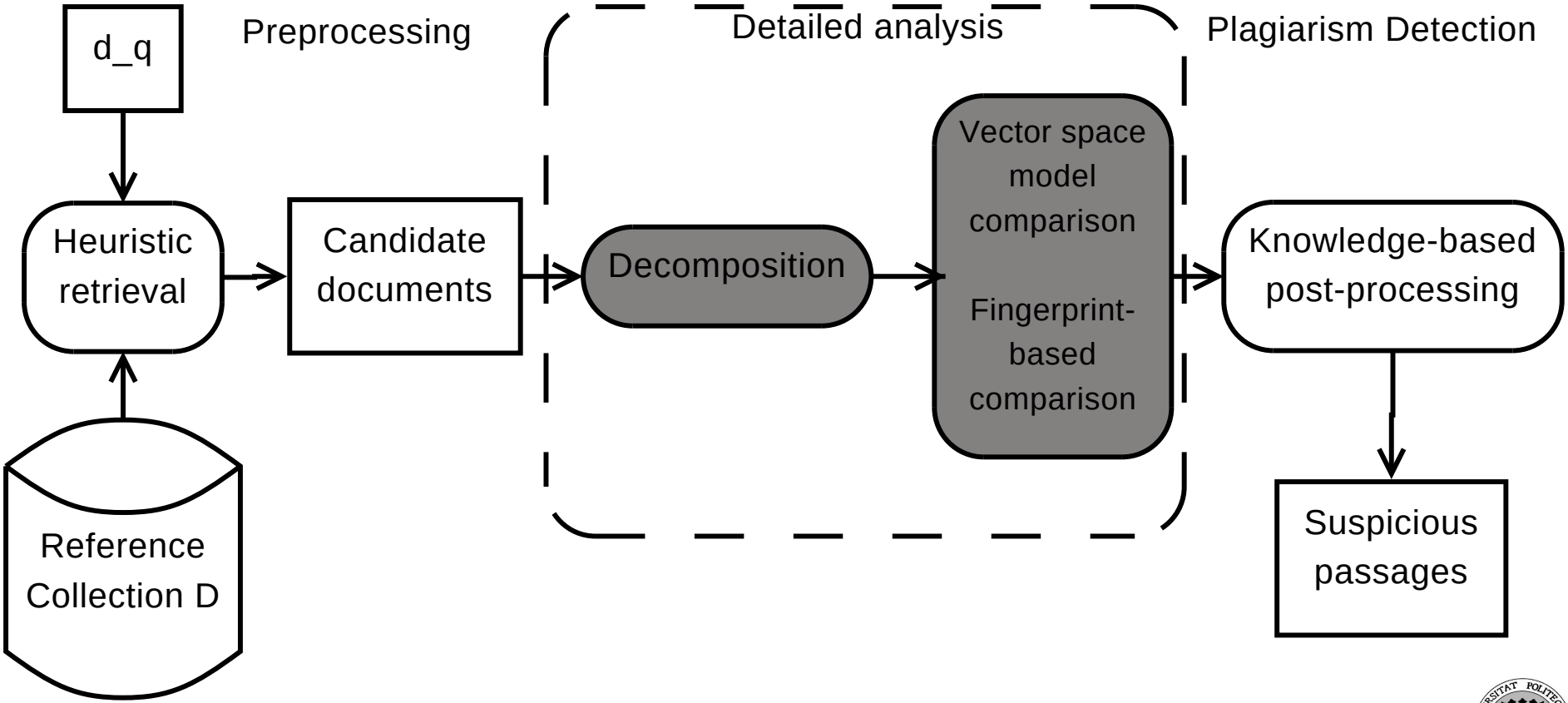
Contributions

Search space reduction

- Retrieval of documents based on query-documents
- Document and keyword-based retrieval are not the same
- Most of the papers on this topic assume that such stage is solved
- It is often called “heuristic source documents retrieval”

[Barrón-Cedeño et al., 2009b, CICLing]

Contributions



Contributions

Cross-Language Approach

- Numerous tools exist, but they do not pay attention to the cross-language case
- Proposal of a method based on statistical machine translation
- Preliminary experiments

- Maybe one of the first papers on this topic
  [Barrón-Cedeño et al., 2008, PAN cl] (also [Pinto et al., 2009, Algorithms])

Outline

Corpora development

- Corpora of real cases of plagiarism have not been published because of ethical reasons
- Nobody wants to be exposed!
- We need to compile/generate corpora

Wikipedia co-derivatives corpus

- Texts written in: en, de, es, hi
- 500 most frequently accessed articles in each language
- For each article 10 revisions are included

http://users.dsic.upv.es/grupos/nle/downloads.html

Monolingual Text Similarity Measures

Comparison of models including:

- Vector space models
- Probabilistic models
- Fingerprinting models

(the most simple seems to be the best: Jaccard Coefficient)

[Barrón-Cedeño et al., 2009a, ICON, in press]

Cross-Language Text Reuse corpus

- It includes parallel texts from the JRC-Acquis corpus
- Comparable texts from Wikipedia are included as well

- Simulation of cross-language text reuse (plagiarism) based on translation and translation+post-edition

http://users.dsic.upv.es/grupos/nle/downloads.html

Cross-Language Similarity Measures

- CL **E**xplicit **S**emantic **A**nalysis [Potthast et al., 2008]
- CL **C**haracter **n**-grams-based **C**omparison [McNamee and Mayfield, 2004]
- CL **A**lignment-based **S**imilarity **A**nalysis [Barrón-Cedeño et al., 2008, Pinto et al., 2009]

- CL-CnC is the best option when the languages are related
- CL-ESA is better on Wikipedia (but Wikipedia pages are far from being plagiarism!)
- CL-ASA is better on exact translations
- CL-ASA can be applied to any pair of languages (related or not)

[Potthast et al., 2009a, LRE, in press]

Development of the PAN-PC-09 corpus

- 41,223 documents including 94,202 cases of artificial plagiarism
- Plagiarism Languages. 90% of the cases are monolingual English plagiarism. The remainder are cross-lingual (from German and Spanish into English).
- Plagiarism Obfuscation.

http://users.dsic.upv.es/grupos/nle/downloads.html

[Potthast et al., 2009b, PAN]

PAN-09@SEPLN: Workshop & Competition

PAN Workshop. Uncovering Plagiarism, Authorship and Social Software Misuse

YAHOO! RESEARCH

http://www.webis.de/pan-09

Lab@CLEF-10: on Plagiarism Detection

Conference on Multilingual and Multimodal Information Access Evaluation

Contact

Master en Inteligencia Artificial, Reconocimiento de Formas e Imagen Digital
http://www.popinformatica.upv.es/iarfid.html

Natural Language Engineering Lab
http://www.dsic.upv.es/grupos/nle/

Alberto Barrón Cedeño
http://www.dsic.upv.es/∼lbarron

Paolo Rosso
http://www.dsic.upv.es/∼prosso

CONACyT

The research work of Alberto Barrón Cedeño is possible thanks to the CONACyT-Mexico support.

References I

Barrón-Cedeño, A., Eiselt, A., and Rosso, P. (2009a).

Monolingual Text Similarity Measures: A comparison of Models over Wikipedia Articles Revisions.
In *Proceedings of the ICON 2009*.

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

On Cross-lingual Plagiarism Analysis using a Statistical Model.
In Stein, Stamatatos, and Koppel, editors, *Proceedings of the ECAI'08 PAN Workshop: Uncovering Plagiarism, Authorship and Social Software Misuse*, pages 9–14, Patras, Greece.

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

Towards the Exploitation of Statistical Language Models for Plagiarism Detection with Reference.
In Stein, Stamatatos, and Koppel, editors, *Proceedings of the ECAI'08 PAN Workshop: Uncovering Plagiarism, Authorship and Social Software Misuse*, pages 15–19, Patras, Greece.

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

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

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

On the Relevance of Search space Reduction in Automatic Plagiarism Detection.
Procesamiento del Lenguaje Natural, 43:141–149.

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

Reducing the Plagiarism Detection Search Space on the Basis of the Kullback-Leibler Distance.
Computational Linguistics and Intelligent Text Processing. Proceedings of the CICLing 2009, LNCS (5449):523–53

References II

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.

Brin, S., Davis, J., and Garcia-Molina, H. (1995).
Copy Detection Mechanisms for Digital Documents.
In *ACM International Conference on Management of Data (SIGMOD 1995)*.

Broder, A. (1997).
On the Resemblance and Containment of Documents.
In *SEQUENCES '97: Proceedings of the Compression and Complexity of Sequences 1997*, page 21, Washington, DC. IEEE Computer Society.

Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Lafferty, J., Mercer, R., and Roossin, P. (1990).
A Statistical Approach to Machine Translation.
Computational Linguistics, 16(2):79–85.

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.

Clough, P. (2000).
Plagiarism in Natural and Programming Languages: an Overview of Current Tools and Technologies.
Research Memoranda: CS-00-05, Department of Computer Science. University of Sheffield, UK.

References III

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.

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

Lyon, C., Malcolm, J., and Dickerson, B. (2001).
Detecting short passages of similar text in large document collections.
In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, pages 118–125, Pennsylvania.

McNamee, P. and Mayfield, J. (2004).
Character n-gram tokenization for european language text retrieval.
Information Retrieval, 7(1–2):73–97.

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.

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

References IV

Potthast, M., Barrón-Cedeño, A., Stein, B., and Prosso, P. (2009a).
Cross-Language Plagiarism Detection.
Languages Resources and Evaluation, Special Issue on Plagiarism and Authorship Analysis.

Potthast, M., Stein, B., and Anderka, M. (2008).
A Wikipedia-Based Multilingual Retrieval Model.
Proceedings of the 30th European Conf. on IR Research (ECIR 2008), LNCS (4956):522–530.

Potthast, M., Stein, B., Eiselt, A., Alberto, B.-C., and Rosso, P. (2009b).
Overview of the 1st International Competition on Plagiarism Detection.
In Stein, Rosso, Stamatatos, Koppel, and Agirre, editors, *SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse (PAN 09)*, pages 1–9. CEUR-WS.org.

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.

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.

Stamatatos, E. (2009).
A Survey of Modern Authorship Attribution Methods.
Journal of the American Society for Information Science and Technology, 60(3):538–556.

References V

Stein, B. (2007).

Principles of hash-based text retrieval.

In Clarke, C., Fuhr, N., Kando, N., Kraaij, W., and de Vries, A., editors, *30th Annual International ACM SIGIR Conference*, pages 527–534, Amsterdam, Netherlands. ACM.

Pinto, Civera, Barrón-Cedeño, Juan, Rosso: *A statistical approach to natural language tasks*, 2009.