Semantic Relation Analysis for Paraphrase Plagiarism Detection on Computing Scientific Texts in Spanish

Paraphrased Text Reuse Detection in Spanish Computing Monolingual Corpora

Meneses-Abad, Abel[1] Madera, Julio[2]
Barron-Cedeño, Alberto[3]

[1]Center for Studies on Informatics at UCLV

[2]Mathematical Dept. at UC

[3]QCRI

2017

About the speaker

10 Years in FreeSoftware, 5 Years in NLP with Python

- Ing. Telecomunicaciones y Electrónica, CUJAE, 2004.
- Miembro Grupo Técnico Nacional de SWL 2005 - 2009.
- Pte. Científico IV Taller Internacional de SWL, Informática 2009.
- Miembro del GUTL del 2009 - a la actualidad.
- Uno de los autores de la Guía Cubana de Migración a SWL de Cuba.
- Jefe del Grupo que logró la migración del 40% de la UCI, de 2005 - 2009.
- Proyectos libres cubanos: SistClon, Infodrez, Sunshine, Shakespeare, QtNLP...
- Desde 2012 realiza su Dr.C. de la Computación, en PLN, Detección de Plagio.

What's it all about?

1. Plagiarism Detection Roadmap
    - The Basic Problem of Plagiarism Detection
    - Previous Works
2. Our Possible Contribution
    - Hypothesis & Explanation
    - Basic Ideas for the Investigation

Outline

1. Plagiarism Detection Roadmap
   - The Basic Problem of Plagiarism Detection
   - Previous Works
2. Our Possible Contribution
   - Hypothesis & Explanation
   - Basic Ideas for the Investigation

What is plagiarism?
Digital Plagiarism Detection Lingeries

- “The act of taking the writings of another person and passing them off as one’s own, generally in violation of copyright laws.”
- Source code plagiarism has been studied before the ’80s. ***[Parker and Hamblen, 1989]*** Internet & Copyleft upgraded that problem!
- Natural languague texts, that represent 85% of Internet available texts, are the most plagiarised archieves nowadays.
- Jon Barrie, founder of Turnitin.com, says that a third of their papers(40 million) have significant levels of plagiarism.
- Some reports suggest that there are more copyright violations among computer science students than in any other academic discipline.***[Barrie and Denning, 2010]***

Paraphrase Plagiarism Recognition

An open issue in computer science?

Examples

Copy & Paste Plagiarism.
Paraphrase Plagiarism
Missing citation.
Data fabrication.
Idea Plagiarism.

Definition

Interest domain of
the investigation:
Paraphrase Plagiarism.

[El Thair et al., 2011][Imran, 2010][Barrón-Cedeño et al., 2010][Kakkonen and Mozgovoy, 2010][Lukashenko et al., 2007] are agree with some of this categories.

Paraphrase Plagiarism Recognition

An open issue in computer science?

Examples

Copy & Paste Plagiarism.
Paraphrase Plagiarism
Missing citation.
Data fabrication.
Idea Plagiarism.

Definition

Interest domain of the investigation: **Paraphrase Plagiarism.**

[El Thair et al., 2011][Imran, 2010][Barrón-Cedeño et al., 2010][Kakkonen and Mozgovoy, 2010][Lukashenko et al., 2007] are agree with some of this categories.

Paraphrase Plagiarism Recognition

An open issue in computer science?

Examples

Copy & Paste Plagiarism.
Paraphrase Plagiarism
Missing citation.
Data fabrication.
Idea Plagiarism.

Definition

Interest domain of the investigation: **Paraphrase Plagiarism.**

It takes into account the existent semantic relations between synonyms, antonyms, polysemic words, etc.

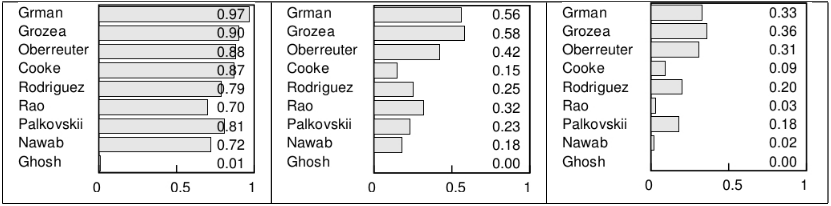
PAN, the biggest world plagiarism detection competition shows that *recall* must improve in paraphrase.

2011 PAN statistics

Results of plagiarism methods competition by types

Non-paraphrase plagiarism detection methods are **over 80%** in the majority of cases.

Recall of Web Paraphrase Plagiarism Detection, by authors.***[Barrón-Cedeño, 2012]***



Other types	**Auto-Low**	**Manual**

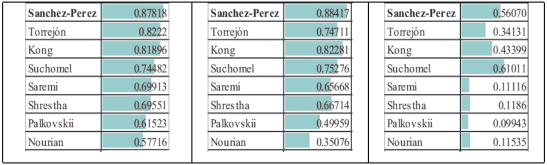
Recall in Auto-High paraphrase plagiarism detection is under 0.1.
Paraphrase plagiarism detection overall is **less** than **40%**.

2014 PAN statistics

Results of plagiarism methods competition by types

Non-paraphrase plagiarism detection methods are **over 80%** [Sanchez-perez, 2014].

Recall of Web Paraphrase Plagiarism Detection, by authors.***[Barrón-Cedeño, 2012]***



	Other types	Random	Summary
Sanchez-Perez	0.87818	0.88417	0.56070
Torrejón	0.8222	0.74711	0.34131
Kong	0.81896	0.82281	0.43399
Suchomel	0.74482	0.75276	0.61011
Saremi	0.69913	0.65668	0.11116
Shrestha	0.69551	0.66714	0.1186
Palkovskii	0.61523	0.49959	0.09943
Nourian	0.57716	0.35076	0.11535

Recall in paraphrase plagiarism detection is over 0.5% only in simulated cases

Real paraphrase plagiarism detection overall is **less** than **20%**.

The gap still open in English! Spanish corpus missing!

Paraphrase Complexity

Learn the way to do it, then understand the difficulties to detect it.

[Barrón-Cedeño et al., 2012] consider the following six paraphrase typologies:

1. ***Morphology* - *based changes:*** Inflections, modal verb & derivational changes.
2. ***Lexicon* - *based changes:*** Spelling, same polarity, synthetic & analitic sustitutions.
3. ***Syntax* - *based changes:*** Negation, ellipsis, etc.
4. ***Discourse* - *based changes:*** Punctuation & format changes, styles alternation, etc.
5. ***Miscellaneous* - *based changes:*** Discourse structures & order changes. Addition & deletion.
6. ***Semantic* - *based changes:*** Imply a different lexicalisation of the same content unit.

Surveillance Plagiarism: Detection Systems

There are two kinds of plagiarism detection systems: Hermetic and Web.

Summarized list from ***[Kakkonen and Mozgovoy, 2010]***

- **WCopyfind**. Hermetic detection system to compare natural language texts.
- **Sherlock**. Hermetic detection system to compare natural language and source code texts.
- **SeeSources.com**. Web detection system that allows users to search the sources of a record length(in words) or a file on the web. It uses classic internet services like Google or Yahoo.
- **TurnitIn**. Hermetic & Web detection system. It is the most used anti-plagiarism software, which produces a report about a text, comparing it with documents comming from Internet and also its own 40 million students papers database.

But are they necessary in national academies?

Slovak Republic Example

The social dimension of plagiarism problem.

Summarized list from ***[Grman and Ravas, 2011]***

- Complex evaluation of thesis and dissertations of all 33 universities in the Slovak Republic
- Language independent solution - documents in Slovak, Czech, Ukrainian, Hungarian and English
- Approx. 80 thousands of thesis and dissertations per year
- 3.4 million documents from internet (06/2011)
- Core detection algorithm is now running on one server only (but parallel processing available)

Other Problems on Plagiarism Detection.

Some things are not plagiarism but affect it.

- Natural languague evolution.
  Words with lexical function can get more meanings through the time, others can just appear. For example: google is nowadays an accepted verb, which means “search on Google”.
- Copyleft copyright
  The copyright analysis must consider the copyleft detection due to the presence - allowed and probable - of long fragments reused not considered plagiarism.
- Common Knowledge
  Anything that is generally known to everyone. Ej: “Independence Day” of a particular contry, or “the saying goes”.

Outline

1. Plagiarism Detection Roadmap
   - The Basic Problem of Plagiarism Detection
   - Previous Works
2. Our Possible Contribution
   - Hypothesis & Explanation
   - Basic Ideas for the Investigation

Plagiarism Detection Methods

Summarized list from ***[El Thair et al., 2011]***

- Grammar Based method.
  Based on grammar, it uses string matching. They are very effective only for copy-paste plagiarism.
- Semantic Based method.
  Based on vector space model, it has low results for partially plagiarized documents.

- **Grammar Semantic hybrid method.**
  Based on specific index structures like tri-grams, n-grams, or hash-based fingerprints. They are suitable for modified texts by rewriting or switching words with the same meaning.

Models on Grammar-Semantic hybrid methods

Who detects better under which circumstances?

- Phrase extraction & semantic analysis.
  *It uses the index coming from IR index, wordnet to detect synonyms, and SCAM measure for detecting candidates.**[Anzelmi et al., 2011]***
- n-grams model.
  *Based on word length secuence encoding, and uses a downstream vector-based n-gram distance measure for candidate selection.**[Basile et al., 2009]***

Models on Grammar-Semantic hybrid methods

Who detects better under which circumstances?

- Term Document Weight matrix & Minimum Weight Overlapping.
  It uses shingles pre-processing, and normalization technique, then uses weights to model de TDW matrix, and finally filtering and MWO to detect duplicate parts.
  [Das et al., 2011]
- Hash-based fingerprint model.
  It incorporates common text shingles in pre-selection, and fingerprints for candidate retrieval.***[Kasprzak et al., 2009]***

Consulted Research Groups

- Some investigated: Slovak group SVOP Ltd, Gensim project Czech Republic Masaryk University.
- Less investigated:
  - **Bauhaus-Universität Weimar's Group**
  - **Linguistic Group Universitat Politècnica de València,**

  Others: Linguistic Engineer UNAM / **European NLP Group** / Spain NLP papers (SEPLN) / **Stanford NLP Group** / other plagiarism works from India, China, Greece
- Future investigations: **Google NLP Group** / Microsoft NLP Group

Plagiarism Detection Process

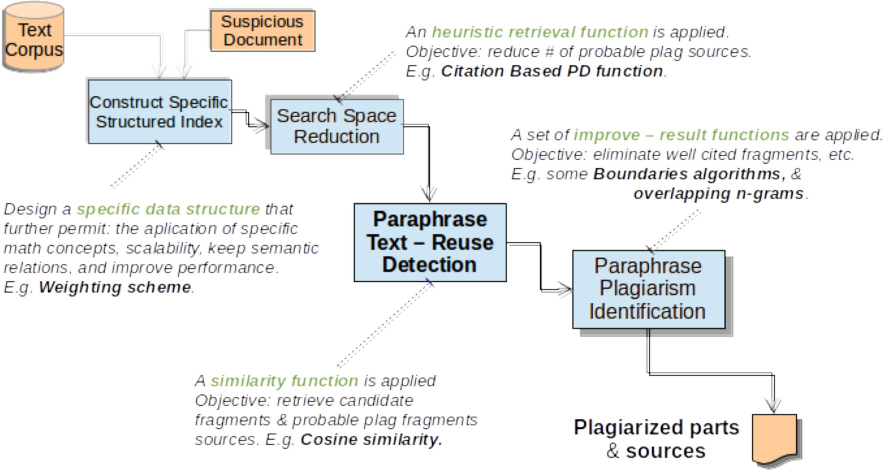
Benno Stein[1] defines that plagiarism detection process has three stages:***[Stein et al., 2007]***

1. ***Heuristic retrieval of potential source documents:*** gives the subset $D^* \in D$ such that $|D^*| \lll |D|$,where D is a collection of reference documents.
2. ***Exhaustive comparison of texts:*** compares the text d_q & $d \in D^*$identifying reused fragments and their potential sources.
3. ***Knowledge - based post processing:*** this phase detects proper citation fragments, common knowledge, etc.

[1] Bauhaus-Universität Weimar Professor, Germany. The most cited scientist in plagiarism investigation by the ACM.

Plagiarism Recognition Process

Authorship diagram by Ph. D. candidate Abel Meneses Abad


Text Corpus
Suspicious Document
Construct Specific Structured Index
Search Space Reduction
Paraphrase Text – Reuse Detection
Paraphrase Plagiarism Identification
Plagiarized parts & sources

An *heuristic retrieval function* is applied.
Objective: reduce # of probable plag sources.
E.g. ***Citation Based PD function***.

A set of *improve – result functions* are applied.
Objective: eliminate well cited fragments, etc.
E.g. some ***Boundaries algorithms***, & ***overlapping n-grams***.

Design a *specific data structure* that further permit: the aplication of specific math concepts, scalability, keep semantic relations, and improve performance.
E.g. ***Weighting scheme***.

A *similarity function* is applied
Objective: retrieve candidate fragments & probable plag fragments sources. E.g. ***Cosine similarity.***

Outline

1. Plagiarism Detection Roadmap
   - The Basic Problem of Plagiarism Detection
   - Previous Works
2. Our Possible Contribution
   - Hypothesis & Explanation
   - Basic Ideas for the Investigation

¿A qué nos enfrentamos?

Spanish semantic analysis doesn't appear on recent investigations.

- La detección del plagio parafrástico es un tema muy actual en las ciencias de la computación.
- Actualmente un número creciente de artículos trata el tema del procesamiento de textos y la detección del plagio parafrástico.
- Existen pocos artículos que refieran resultados exhaustivos de la detección de plagio monolingue para el idioma español.
- Frecuentemente la reducción del problema en NLP se hace mediante la selección de corpus temáticos.
  Ej. procesamiento de textos médicos.

Semantic Relations in Paraphrase

How can affect semantic relations paraphrase changes?

Our considerations from ***[Barrón-Cedeño et al., 2012]*** paraphrase typologies:

1. ***Morphology - based changes:*** Inflections, modal verb & derivational changes.
2. ***Lexicon - based changes:*** Spelling, same polarity, synthetic & analitic sustitutions.
3. ***Syntax - based changes:*** Negation, ellipsis, etc.
4. ***Discourse - based changes:*** Punctuation & format changes, styles alternation, etc.
5. ***Miscellaneous - based changes:*** Discourse structures & order changes. Addition & deletion.
6. ***Semantic - based changes:*** Imply a different lexicalisation of the same content unit.

These typologies are closely related with "Linguistic Semantic Relations".

Hypothesis

Spanish semantic analysis can have a better performance.

La utilización de *estructuras de datos* y *algoritmos de similitud* novedosos que permitan guardar y recuperar las *relaciones semánticas* existentes en el *idioma español*, podría mostrar mejores resultados en la detección de *plagio parafrástico*, mejorando su eficiencia al introducir modificaciones en el contexto de los *textos científicos sobre computación*.

Explanation

- La experimentación de soluciones basadas en modelos computacionales que permitan conservar información sobre las relaciones semánticas, así como su rediseño, optimización y escalabilidad en la detección de plagio parafrástico en idioma español, es de vital importancia para mejorar el desarrollo de los procesos de enseñanza, fundamentalmente en las universidades y centros que generan y resguardan patrimonio documental creciente en el tiempo.
- La realización de este proceso de forma manual es imposible, la existencia de un proceso automático que permita mejorar la redacción y revisión de los textos científicos constituye una ventaja para el mejoramiento de los servicios documentales.

Outline

1. Plagiarism Detection Roadmap
   - The Basic Problem of Plagiarism Detection
   - Previous Works
2. Our Possible Contribution
   - Hypothesis & Explanation
   - Basic Ideas for the Investigation

Specific Objectives

Investigation Roadmap & Milestones

- Diseñar variantes de estructuras de datos y algoritmos que comprendan las relaciones semánticas del idioma español, reutilizando ideas existentes para otros lenguajes naturales.
- Valorar estadísticamente los resultados de la aplicación de las estructuras y algoritmos diseñados para la detección de plagio parafrástico comparándolos con otros reportados en la literatura para idioma inglés y español.
- Aplicar la mejor variante diseñada al procesamiento de textos científicos de computación en español de las universidades de Camagüey y Granma mediante el desarrollo de la plataforma Sunshine de Gestión Documental.

What kind of software is it involved?

Pretended practical goal: an hermetic plagiarism detection system.

A solution like **WCopyfind.**
Only an hermetic detection system that compares computing scientific texts from essays databases of Cuban universities.

Identify paraphrase plagiarism on the web is not a purpose of this investigation.

Conclusions

- Cuba has had few works on this area of knowledge in the last years.
- Plagiarism detection software are not optimized for environments with low hardware tech.
- Spanish plagiarism detection is a rare topic on the actual NLP investigations.

Conclusions

Cuban educational technology urgently needs a plagiarism detection service.

- Developping Sunshine [2] can be created a solid background for this investigation and others related in Cuba.
- By applying other Cuban doctoral results like clustering and EDA optimized algorithms, it can be obtained a more original and complete solution for Cuban text mining investigations.

- **Outlook**
  - The integration of Spanish natural languague results of Cuban investigations into digital Spanish dictionaries & thesaurus has never been done.

[2] A cuban project of Document Management System, that use a python toolkit and resources.

For Further Reading I

Daniele Anzelmi, Domenico Carlone, Fabio Rizzello, Robert Thomsen, and D M Akbar Hussain. Plagiarism Detection Based on SCAM Algorithm. In *International MultiConference of Engineers and Computer Scientists*, volume I, page 6, Hong Kong, 2011. ISBN 9789881821034.

John Barrie and Peter J Denning. Cheating in Computer Science. *Ubiquity*, (October):1–5, 2010.

Alberto Barrón-Cedeño. *On the Mono- and Cross-Language Detection of Text Re-Use and Plagiarism*. Doctoral thesis, Universidad Politécnica de Valencia, 2012.

For Further Reading II

Alberto Barrón-Cedeño, Marta Vila, and Paolo Rosso. Detección automática de plagio: de la copia exacta a la paráfrasis. In Euphonia Ediciones SL, editor, *Panorama actual de la lingüística forense en el ámbito legal y policial: Teoría y práctica. Madrid.*, number 2007, pages 1–19, Madrid, 2010.

Alberto Barrón-Cedeño, Marta Vila, M. Antonia Martí, and Paolo Rosso. Plagiarism meets Paraphrasing: Insights for the Next Generation in Automatic Plagiarism Detection. *Computational Linguistics*, (October), 2012. doi: 10.1162/COLI.

Chiara Basile, Dario Benedetto, Emanuele Caglioti, Giampaolo Cristadoro, and Mirko Degli Esposti. A plagiarism detection procedure in three steps: selection, matches and squares. In *PAN'2009*, pages 19–23, 2009.

For Further Reading III

Shine N Das, Midhun Mathew, and Pramod K Vijayaraghavan. An Efficient Approach for Finding Near Duplicate Web pages using Minimum Weight Overlapping Method. 1(2):187–195, 2011.

Asim M El Thair, Hussam M. Dahwa, and Vaclav Snasel. Survey of Plagiarism Detection Methods. In *Fifth Asia Modellingg Symposium*, pages 39–42, 2011. ISBN 9780769544144. doi: 10.1109/AMS.2011.19.

Jan Grman and Rudolf Ravas. Improved implementation for finding text similarities in large collections of data. In *PAN'2011*, page 18, 2011.

Naveed Imran. Electronic media, creativity and plagiarism. *ACM SIGCAS Computers and Society*, 40(4):25–44, December 2010. ISSN 00952737. doi: 10.1145/1929609.1929613. URL http://portal.acm.org/citation.cfm?doid=1929609.1929613.

For Further Reading IV

Tuomo Kakkonen and Maxim Mozgovoy. Hermetic and web plagiarism detection systems for student essays an evaluation of the state-of-the-art. 42(2):135–159, 2010. doi: 10.2190/EC.42.2.a.

Jan Kasprzak, Michal Brandejs, and Miroslav Kripac. Finding Plagiarism by Evaluating Document Similarities. In *PAN'2009*, pages 24–28, 2009.

Romans Lukashenko, Vita Graudina, and Janis Grundspenkis. Computer-Based Plagiarism Detection Methods and Tools: An Overview. In *CompSysTech'07 International Conference on Computer Systems and Technologies*, pages 1–6. ACM Press, 2007. ISBN 9789549641509.

For Further Reading V

Alan Parker and James O. Hamblen. Computer Algorithms for Plagiarism Detection. *IEEE Transactions on Education*, 32(2): 94–99, 1989. doi: 0018-9359/89/0500-0094.

Miguel A Sanchez-perez. A Winning Approach to Text Alignment for Text Reuse Detection at PAN 2014. In *PAN'2014*, 2014.

Benno Stein, Sven Meyer, and Martin Potthast. Strategies for Retrieving Plagiarized Documents. In *SIGIR'07*, pages 5–6, 2007. ISBN 9781595935977. doi: 978-1-59593-597-7/07/0007.