Paraphrasing

Potential Applications for Plagiarism Detection

Marta Vila

CLiC – Universitat de Barcelona

WIQE 2010 Valencia

September 13, 2010

Outline

1. Paraphrasing and Plagiarism Detection
2. Typologies
   - Paraphrasing Complexity
   - State of the Art
   - A Paraphrasing Typology
3. Corpora
   - State of the Art
   - CoCo Interface
4. NLP Approaches
   - State of the Art
5. The WRPA System
   - Presentation
   - Methodology
   - Evaluation and Results
   - Conclusions and Future Work

Outline

1. Paraphrasing and Plagiarism Detection
2. Typologies
   - Paraphrasing Complexity
   - State of the Art
   - A Paraphrasing Typology
3. Corpora
   - State of the Art
   - CoCo Interface
4. NLP Approaches
   - State of the Art
5. The WRPA System
   - Presentation
   - Methodology
   - Evaluation and Results
   - Conclusions and Future Work

Why paraphrasing?

Paraphrasing stands for (approximate) **sameness of meaning** between **different wordings**.

- He announced that the company wouldn't participate
  He stated that the company wouldn't participate
- My father built the house
  The house was built by my father

Why paraphrasing?

Paraphrasing stands for (approximate) **sameness of meaning** between **different wordings**.

- He announced that the company wouldn't participate
  He stated that the company wouldn't participate
- My father built the house
  The house was built by my father

Why is paraphrasing relevant for plagiarism?

Why paraphrasing?

Paraphrasing stands for (approximate) **sameness of meaning** between **different wordings**.

- He announced that the company wouldn't participate
  He stated that the company wouldn't participate
- My father built the house
  The house was built by my father

Why is paraphrasing relevant for plagiarism?

Paraphrasing is a linguistic ability used in plagiarism.
To plagiarize involves, on many occasions, paraphrasing.

- Hamlet's pretense of madness
  Hamlet adopts a pretense of madness[1]

[1] http://www.princeton.edu/pr/pub/integrity/08/plagiarism/

What is relevant in paraphrasing?

- Paraphrasing typologies
  - → A paraphrasing typology is also a plagiarism phenomena typology

What is relevant in paraphrasing?

- Paraphrasing typologies
  - → A paraphrasing typology is also a plagiarism phenomena typology
- Paraphrasing corpora
- NLP approaches to paraphrasing
  - → A number of techniques and corpora can also be useful in plagiarism detection

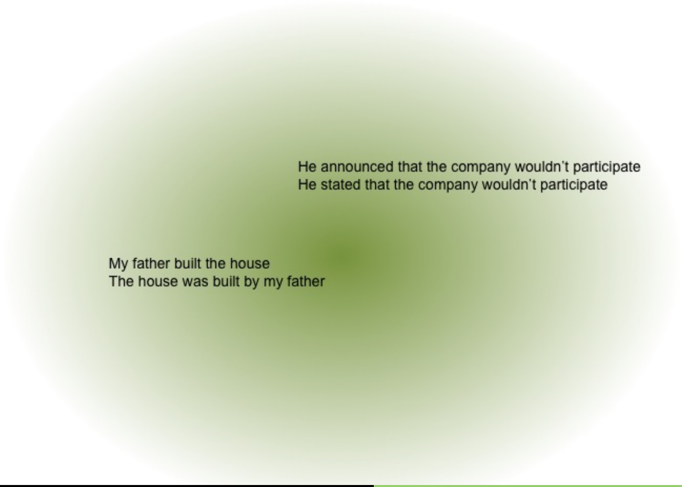
Outline

1. Paraphrasing and Plagiarism Detection
2. Typologies
   - Paraphrasing Complexity
   - State of the Art
   - A Paraphrasing Typology
3. Corpora
   - State of the Art
   - CoCo Interface
4. NLP Approaches
   - State of the Art
5. The WRPA System
   - Presentation
   - Methodology
   - Evaluation and Results
   - Conclusions and Future Work

Outline

1. Paraphrasing and Plagiarism Detection
2. Typologies
   - Paraphrasing Complexity
   - State of the Art
   - A Paraphrasing Typology
3. Corpora
   - State of the Art
   - CoCo Interface
4. NLP Approaches
   - State of the Art
5. The WRPA System
   - Presentation
   - Methodology
   - Evaluation and Results
   - Conclusions and Future Work

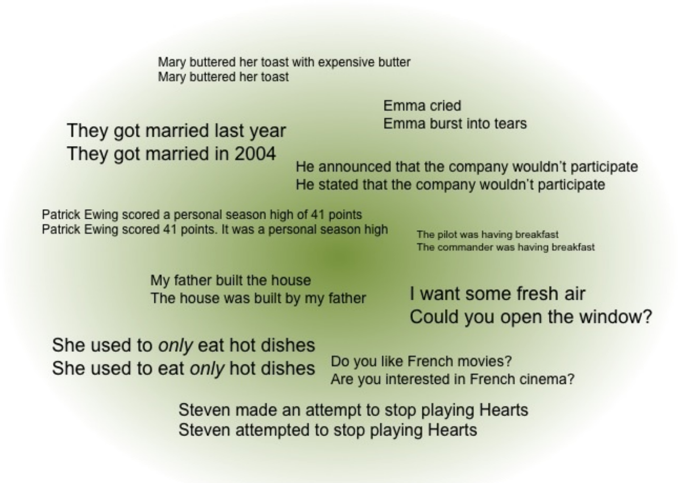
Paraphrasing Complexity



He announced that the company wouldn’t participate
He stated that the company wouldn’t participate

My father built the house
The house was built by my father

Paraphrasing Complexity



Mary buttered her toast with expensive butter
Mary buttered her toast

Emma cried
Emma burst into tears

They got married last year
They got married in 2004

He announced that the company wouldn't participate
He stated that the company wouldn't participate

Patrick Ewing scored a personal season high of 41 points
Patrick Ewing scored 41 points. It was a personal season high

The pilot was having breakfast
The commander was having breakfast

My father built the house
The house was built by my father

I want some fresh air
Could you open the window?

She used to *only* eat hot dishes
She used to eat *only* hot dishes

Do you like French movies?
Are you interested in French cinema?

Steven made an attempt to stop playing Hearts
Steven attempted to stop playing Hearts

Paraphrasing Complexity

Let’s put it in order!

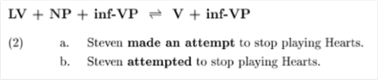
Outline

1. Paraphrasing and Plagiarism Detection
2. Typologies
   - Paraphrasing Complexity
   - State of the Art
   - A Paraphrasing Typology
3. Corpora
   - State of the Art
   - CoCo Interface
4. NLP Approaches
   - State of the Art
5. The WRPA System
   - Presentation
   - Methodology
   - Evaluation and Results
   - Conclusions and Future Work

State of the Art

Paraphrasing typologies:

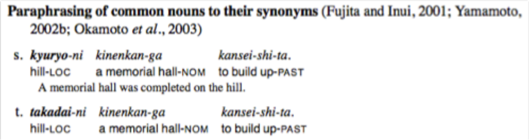
- Dras (1999): syntactic paraphrases



LV + NP + inf-VP ⇌ V + inf-VP

(2) a. Steven **made an attempt** to stop playing Hearts.
b. Steven **attempted** to stop playing Hearts.

- Fuijta (2005): lexical and structural paraphrases



Paraphrasing of common nouns to their synonyms (Fujita and Inui, 2001; Yamamoto, 2002b; Okamoto *et al*., 2003)

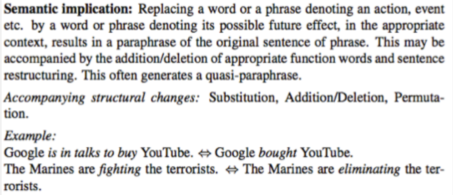
s. ***kyuryo**-ni* *kinenkan-ga* *kansei-shi-ta*.
hill-LOC a memorial hall-NOM to build up-PAST
A memorial hall was completed on the hill.

t. ***takadai**-ni* *kinenkan-ga* *kansei-shi-ta*.
hill-LOC a memorial hall-NOM to build up-PAST

State of the Art

Paraphrasing typologies:

- Bhagat (2009): lexical and structural paraphrases



Semantic implication: Replacing a word or a phrase denoting an action, event etc. by a word or phrase denoting its possible future effect, in the appropriate context, results in a paraphrase of the original sentence of phrase. This may be accompanied by the addition/deletion of appropriate function words and sentence restructuring. This often generates a quasi-paraphrase.

Accompanying structural changes: Substitution, Addition/Deletion, Permutation.

Example:
Google *is in talks to buy* YouTube. ⇔ Google *bought* YouTube.
The Marines are *fighting* the terrorists. ⇔ The Marines are *eliminating* the terrorists.

- Žolkovskij and Mel'čuk (1965, 1966, 1967) and Milićević (2007): syntactic and semantic paraphrases

  a republic/a republican state:

$$C_0 \Leftrightarrow \text{Gener}(C_0) \quad \xrightarrow{\text{ATTR}} A_0/\text{Adv}_1(C_0)$$

State of the Art

Problems:

- They do not cover the paraphrasing phenomenon as a whole
- Tied to a specific linguistic theory and formalism
  - → Meaning-Text Theory in Žolkovskij and Mel'čuk, and Milićević

Outline

1. Paraphrasing and Plagiarism Detection
2. Typologies
   - Paraphrasing Complexity
   - State of the Art
   - A Paraphrasing Typology
3. Corpora
   - State of the Art
   - CoCo Interface
4. NLP Approaches
   - State of the Art
5. The WRPA System
   - Presentation
   - Methodology
   - Evaluation and Results
   - Conclusions and Future Work

A Paraphrasing Typology

- Substitution
  - → He **announced** that the company wouldn't participate
    He **stated** that the company wouldn't participate
- Deletion
  - → Mary opened the **bottle of** wine
    Mary opened the wine
- Transformation
  - → My father built the house
    The house was built by my father
- Splitting
  - → Patrick Ewing scored a personal season high of 41 points
    Patrick Ewing scored 41 points. It was a personal season high
- Change of order
  - → She used to **only** eat hot dishes
    She used to eat **only** hot dishes[1]

[1] Examples extracted from Bhagat (2009), Dras (1999), Fuijta (2005), Pustejovsky (1995) and ourselves.

A Paraphrasing Typology

- Substitution
  - → He **announced** that the company wouldn't participate
    He **stated** that the company wouldn't participate
- Deletion (Addition)
  - → Mary opened the **bottle of** wine
    Mary opened the wine
- Transformation
  - → My father built the house
    The house was built by my father
- Splitting (Combining)
  - → Patrick Ewing scored a personal season high of 41 points
    Patrick Ewing scored 41 points. It was a personal season high
- Change of order
  - → She used to **only** eat hot dishes
    She used to eat **only** hot dishes

A Paraphrasing Typology

- Substitution
  - → He **announced** that the company wouldn't participate
    He **stated** that the company wouldn't participate
- Deletion
  - → Mary opened the **bottle of** wine
    Mary opened the wine
- Transformation
  - → My father built the house
    The house was built by my father
- Splitting
  - → Patrick Ewing scored a personal season high of 41 points
    Patrick Ewing scored 41 points. It was a personal season high
- Change of order
  - → She used to **only** eat hot dishes
    She used to eat **only** hot dishes

Usually combined!

A Paraphrasing Typology

Substitution

→ He **announced** that the company wouldn't participate
He **stated** that the company wouldn't participate

- Deletion
  → Mary opened the **bottle of** wine
  Mary opened the wine
- Transformation
  → My father built the house
  The house was built by my father
- Splitting
  → Patrick Ewing scored a personal season high of 41 points
  Patrick Ewing scored 41 points. It was a personal season high
- Change of order
  → She used to **only** eat hot dishes
  She used to eat **only** hot dishes

Substitution

- Synonymy

  → He **announced** that the company wouldn't participate
  He **stated** that the company wouldn't participate
  ✓WordNet, dictionary of synonyms, thesaurus

Substitution

- Synonymy

  → He **announced** that the company wouldn't participate
  He **stated** that the company wouldn't participate
  ✓WordNet, dictionary of synonyms, thesaurus

  → **Since** it was sunny yesterday, the laundry dried well
  As it was sunny yesterday, the laundry dried well
  ✓Dictionary of synonyms, thesaurus

Substitution

- Synonymy
  - → He **announced** that the company wouldn't participate
    He **stated** that the company wouldn't participate
    ✓WordNet, dictionary of synonyms, thesaurus
  - → **Since** it was sunny yesterday, the laundry dried well
    As it was sunny yesterday, the laundry dried well
    ✓Dictionary of synonyms, thesaurus
- Generalization
  - → I have been studying the reproduction of **cats** for ten years
    I have been studying the reproduction of **felines** for ten years
    ✓WordNet, taxonomies

Substitution

- Synonymy

  → He **announced** that the company wouldn't participate
  He **stated** that the company wouldn't participate
  ✓WordNet, dictionary of synonyms, thesaurus

  → **Since** it was sunny yesterday, the laundry dried well
  As it was sunny yesterday, the laundry dried well
  ✓Dictionary of synonyms, thesaurus

- Generalization

  → I have been studying the reproduction of **cats** for ten years
  I have been studying the reproduction of **felines** for ten years
  ✓WordNet, taxonomies

  → Mary **has worn** soft contact lenses since college
  Mary **has used** soft contact lenses since college
  ? WordNet, taxonomies

Substitution

- Antonymy
  - → The neighboring town is **poorer** in forest resources than our town

    Our town is **richer** in forest resources that the neighboring town

    ✓WordNet, dictionary of antonyms, thesaurus

    ✓Matching

Substitution

- Antynomy

→ The neighboring town is **poorer** in forest resources than our town

Our town is **richer** in forest resources that the neighboring town

✔WordNet, dictionary of antonyms, thesaurus

✔Matching

→ I **lost interest** in the endeavor

I **developed disinterest** in the endeavor

? WordNet, dictionary of antonyms, thesaurus

Substitution

- Actant-Action Substitution

  → I dislike rash **drivers**
  I dislike rash **driving**
  ✓Lemmatizer

Substitution

- Actant-Action Substitution

  → I dislike rash **drivers**
  I dislike rash **driving**
  ✓Lemmatizer

  → Mary is John's **student**
  John **teaches** Mary
  ✓Argument structure

Use of the GL lexical entries (Pustejovsky, 2005) for paraphrasing

$$\begin{bmatrix} \textbf{teach} \\ \text{ARGSTR} = \begin{bmatrix} \text{ARG}_1 = \boxed{1}\,[\textbf{teacher}] \\ \text{ARG}_2 = \boxed{2}\,[\textbf{student}] \end{bmatrix} \\ \text{QUALIA} = \left[\text{AGENT} = \textbf{teach\_act}\,(e_1, \boxed{1}, \boxed{2})\right] \end{bmatrix}$$

Substitution

- Word-Definition Substitution

  → Temporary space for rubble and **scrap** wood
  Temporary space for rubble and wood **that became unnecessary**
  ✓MRD, WordNet

A Paraphrasing Typology

- Substitution

  → He **announced** that the company wouldn't participate
  He **stated** that the company wouldn't participate

Deletion

→ Mary opened the **bottle of** wine
Mary opened the wine

- Transformation

  → My father built the house
  The house was built by my father

- Splitting

  → Patrick Ewing scored a personal season high of 41 points
  Patrick Ewing scored 41 points. It was a personal season high

- Change of order

  → She used to **only** eat hot dishes
  She used to eat **only** hot dishes

Deletion

- Non-propositional content deletion

  → Steven **made an attempt** to stop playing Hearts
  Steven **attempted** to stop paying Hearts
  ✓List of light verbs

Deletion

- Non-propositional content deletion
  - → Steven **made an attempt** to stop playing Hearts
    Steven **attempted** to stop paying Hearts
    ✓List of light verbs
- Argument deletion
  - → **My father** built the house
    The house was built
    ✓Argument structure, diathesis alternations

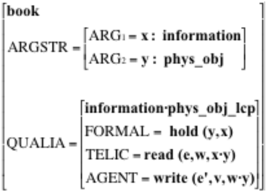
Deletion

- Non-propositional content deletion
  - → Steven **made an attempt** to stop playing Hearts
    Steven **attempted** to stop paying Hearts
    ✓List of light verbs
- Argument deletion
  - → **My father** built the house
    The house was built
    ✓Argument structure, diathesis alternations
- Adjunct deletion
  - → John ran home **at noon**
    John ran home
    ? Argument structure, diathesis alternations

Deletion

- Deletion recoverable by coercion

  → John began **reading** a book
  John began a book
  ✓Qualia



$$\begin{bmatrix} \textbf{book} \\ \text{ARGSTR} = \begin{bmatrix} \text{ARG}_1 = \textbf{x}:\ \textbf{information} \\ \text{ARG}_2 = \textbf{y}:\ \textbf{phys\_obj} \end{bmatrix} \\ \text{QUALIA} = \begin{bmatrix} \textbf{information} \cdot \textbf{phys\_obj\_lcp} \\ \text{FORMAL} = \textbf{hold (y,x)} \\ \text{TELIC} = \textbf{read (e,w,x} \cdot \textbf{y)} \\ \text{AGENT} = \textbf{write (e',v,w} \cdot \textbf{y)} \end{bmatrix} \end{bmatrix}$$

A Paraphrasing Typology

- Substitution

  → He **announced** that the company wouldn't participate
  He **stated** that the company wouldn't participate

- Deletion

  → Mary opened the **bottle of** wine
  Mary opened the wine

Transformation

→ My father built the house
The house was built by my father

- Splitting

  → Patrick Ewing scored a personal season high of 41 points
  Patrick Ewing scored 41 points. It was a personal season high

- Change of order

  → She used to **only** eat hot dishes
  She used to eat **only** hot dishes

Transformation

- Syntactic transformation
  - → My father built the house
    The house was built by my father
  - → The laundry sways in the breeze
    The breeze makes the laundry sway
    ✓Diathesis alternations, Dras (1999)

A Paraphrasing Typology

- Substitution

  → He **announced** that the company wouldn't participate
  He **stated** that the company wouldn't participate

- Deletion

  → Mary opened the **bottle of** wine
  Mary opened the wine

- Transformation

  → My father built the house
  The house was built by my father

Splitting

→ Patrick Ewing scored a personal season high of 41 points
Patrick Ewing scored 41 points. It was a personal season high

- Change of order

  → She used to **only** eat hot dishes
  She used to eat **only** hot dishes

Splitting

- Splitting
  - → Patrick Ewing scored a personal season high of 41 points
    Patrick Ewing scored 41 points. It was a personal season high
    ✓Matching

A Paraphrasing Typology

- Substitution

  → He **announced** that the company wouldn't participate
  He **stated** that the company wouldn't participate

- Deletion

  → Mary opened the **bottle of** wine
  Mary opened the wine

- Transformation

  → My father built the house
  The house was built by my father

- Splitting

  → Patrick Ewing scored a personal season high of 41 points
  Patrick Ewing scored 41 points. It was a personal season high

Change of order

→ She used to **only** eat hot dishes
She used to eat **only** hot dishes

Change of order

- Change of order
  - → She used to **only** eat hot dishes
    She used to eat **only** hot dishes
  - → The student copied the critical diagrams **before returning the book**
    Before returning the book, the student copied the critical diagrams
    ✓Matching

Outline

1. Paraphrasing and Plagiarism Detection
2. Typologies
   - Paraphrasing Complexity
   - State of the Art
   - A Paraphrasing Typology
3. Corpora
   - State of the Art
   - CoCo Interface
4. NLP Approaches
   - State of the Art
5. The WRPA System
   - Presentation
   - Methodology
   - Evaluation and Results
   - Conclusions and Future Work

Outline

1. Paraphrasing and Plagiarism Detection
2. Typologies
   - Paraphrasing Complexity
   - State of the Art
   - A Paraphrasing Typology
3. Corpora
   - State of the Art
   - CoCo Interface
4. NLP Approaches
   - State of the Art
5. The WRPA System
   - Presentation
   - Methodology
   - Evaluation and Results
   - Conclusions and Future Work

Corpora

Corpora used for paraphrasing treatment:

- Parallel corpora (multiple parallel translations, source)

  Barzilay and McKeown (2001)
- Comparable corpora (newspaper articles, same event)

  Barzilay, McKeown and Elhadad (1999)
- Bilingual parallel corpora (source – translation)

  Zhao et al. (2009)
- Monolingual corpus

  Bhagat and Ravichandran (2008)
- Wikipedia

  Vila, Rodríguez and Martí (2010)
- Web

  Dolan, Quirk, Brockett (2004)

Paraphrasing corpora:

- Microsoft Research Paraphrase Corpus

  Dolan, Brockett and Quirk (2005)

- Barzilay and Lee (2003)

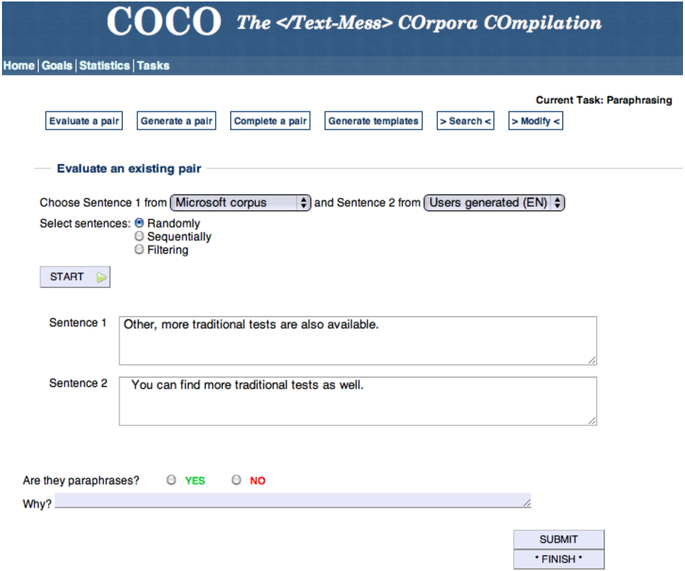
  `http://www.cs.cornell.edu/Info/Projects/NLP/statpar.html`

- CoCo interface

  España et al. (2009)

Outline

1. Paraphrasing and Plagiarism Detection
2. Typologies
   - Paraphrasing Complexity
   - State of the Art
   - A Paraphrasing Typology
3. Corpora
   - State of the Art
   - CoCo Interface
4. NLP Approaches
   - State of the Art
5. The WRPA System
   - Presentation
   - Methodology
   - Evaluation and Results
   - Conclusions and Future Work



Current Task: Paraphrasing

Evaluate an existing pair

Choose Sentence 1 from Microsoft corpus and Sentence 2 from Users generated (EN)

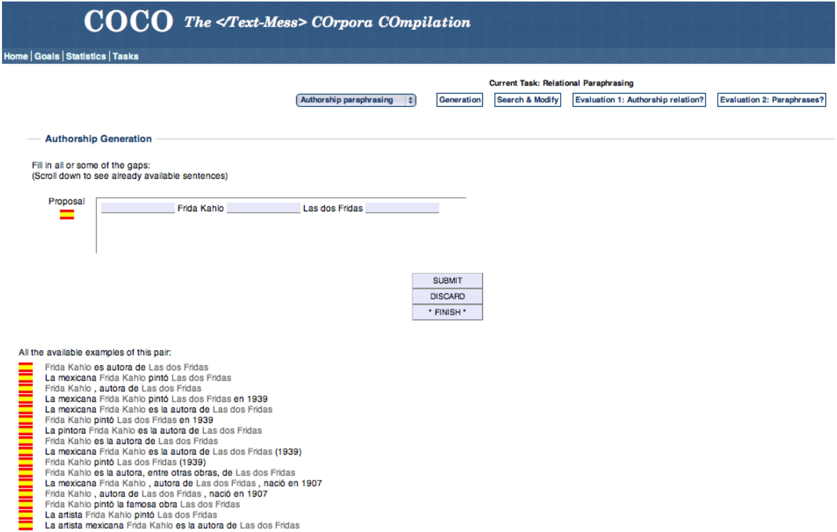
Select sentences: ● Randomly
○ Sequentially
○ Filtering

Sentence 1: Other, more traditional tests are also available.

Sentence 2: You can find more traditional tests as well.

Are they paraphrases? ○ YES ○ NO

Why?


COCO The </Text-Mess> COrpora COmpilation
Home | Goals | Statistics | Tasks
Authorship paraphrasing
Generation
Search & Modify
Evaluation 1: Authorship relation?
Evaluation 2: Paraphrases?
SUBMIT
DISCARD
* FINISH *

Current Task: Relational Paraphrasing

Authorship Generation

Fill in all or some of the gaps:
(Scroll down to see already available sentences)

Proposal ______ Frida Kahlo ______ Las dos Fridas ______

All the available examples of this pair:

Frida Kahlo es autora de Las dos Fridas
La mexicana Frida Kahlo pintó Las dos Fridas
Frida Kahlo , autora de Las dos Fridas
La mexicana Frida Kahlo pintó Las dos Fridas en 1939
La mexicana Frida Kahlo es la autora de Las dos Fridas
Frida Kahlo pintó Las dos Fridas en 1939
La pintora Frida Kahlo es la autora de Las dos Fridas
Frida Kahlo es la autora de Las dos Fridas
La mexicana Frida Kahlo es la autora de Las dos Fridas (1939)
Frida Kahlo pintó Las dos Fridas (1939)
Frida Kahlo es la autora, entre otras obras, de Las dos Fridas
La mexicana Frida Kahlo , autora de Las dos Fridas , nació en 1907
Frida Kahlo , autora de Las dos Fridas , nació en 1907
Frida Kahlo pintó la famosa obra Las dos Fridas
La artista Frida Kahlo pintó Las dos Fridas
La artista mexicana Frida Kahlo es la autora de Las dos Fridas

Outline

1. Paraphrasing and Plagiarism Detection
2. Typologies
   - Paraphrasing Complexity
   - State of the Art
   - A Paraphrasing Typology
3. Corpora
   - State of the Art
   - CoCo Interface
4. NLP Approaches
   - State of the Art
5. The WRPA System
   - Presentation
   - Methodology
   - Evaluation and Results
   - Conclusions and Future Work

Outline

1. Paraphrasing and Plagiarism Detection
2. Typologies
   - Paraphrasing Complexity
   - State of the Art
   - A Paraphrasing Typology
3. Corpora
   - State of the Art
   - CoCo Interface
4. NLP Approaches
   - State of the Art
5. The WRPA System
   - Presentation
   - Methodology
   - Evaluation and Results
   - Conclusions and Future Work

State of the Art

Distributional Hypothesis (Harris, 1954)

Words that occur in the same contexts tend to have similar meanings.

The students solved the problem
The students found a solution to the problem

Bhagat and Ravichandran (2008)

Extended Distributional Hypothesis (Lin and Pantel, 2001)

Paths that link the same sets of words tend to have similar meanings.

The students ←N:subj:V←solved→V:obj:N→the problem
The students ← N:subj:V←find→V:obj:N→solution→N:to:N→the problem

State of the Art

Matching

Strings that share a high number of units tend to have similar meanings.

- Bag of words (NEs)
- N-grams

DATE: NUM1 are killed and around NUM2 injured when suicide bomber blows up his explosive-packed belt at X1 in X2.

palestinian suicide bomber blew himself up at X1 in X2 DATE, killing NUM1 and wounding NUM2 police said.

Barzilay and Lee (2003)

State of the Art

Edit distance (Levenshtein Distance)

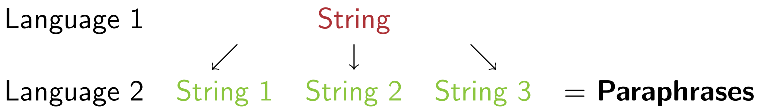
Strings separated by a small edit distance tend to have similar meanings.

Dolan, Quirk and Brockett (2004)

State of the Art

Multiple translation

Strings resulting from the translation of another string in another language are paraphrases.



Language 1 String

Language 2 String 1 String 2 String 3 = **Paraphrases**

Barzilay and McKeown (2001)

Zhao et al. (2009)

State of the Art

Rule application

To check if strings satisfy a set of (manually) created paraphrasing rules.

Head omission: group of students/students
Ordering of sentence components: Tuesday they met.../They met ... Tuesday

Barzilay, McKeown and Elhadad (1999)

Outline

1. Paraphrasing and Plagiarism Detection
2. Typologies
   - Paraphrasing Complexity
   - State of the Art
   - A Paraphrasing Typology
3. Corpora
   - State of the Art
   - CoCo Interface
4. NLP Approaches
   - State of the Art
5. The WRPA System
   - Presentation
   - Methodology
   - Evaluation and Results
   - Conclusions and Future Work

Outline

1. Paraphrasing and Plagiarism Detection
2. Typologies
   - Paraphrasing Complexity
   - State of the Art
   - A Paraphrasing Typology
3. Corpora
   - State of the Art
   - CoCo Interface
4. NLP Approaches
   - State of the Art
5. The WRPA System
   - Presentation
   - Methodology
   - Evaluation and Results
   - Conclusions and Future Work

WRPA
A System for Relational Paraphrase Acquisition from Wikipedia

Vila, Rodríguez and Martí (2010)

Relational Paraphrases

Paraphrases expressing a relation between two entities <*Source*, Target>.

- Cervantes **wrote** El Quijote/El Quijote **by** Cervantes
- Joan Ponç **was born in** Barcelona/Barcelona, Joan Ponç**'s home town**

Hypothesis

Distributional Hypothesis (Harris, 1954)

Words that occur in the same contexts tend to have similar meanings.

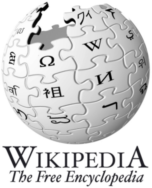
The students	**solved**	the problem
The students	**found a solution to**	the problem

Our hypothesis:

same *Sources* and *Targets* ↔ paraphrase candidates

	AUTHOR	**wrote**	WORK
The designer of	WORK	**was**	AUTHOR
	WORK	**was created by**	AUTHOR
	AUTHOR	**, inventor of**	WORK
	WORK	**by**	AUTHOR

WRPA–Relational Paraphrase Aquisition from Wikipedia



→

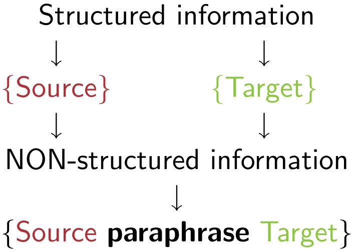
Wu y Weld (2007, 2010)
Wu, Hoffmann y Weld (2008)

WRPA	Others
Complex relations	Simple relations
+ Other sections	+ Infoboxes
Paraphrasing	Information Extraction

Outline

1. Paraphrasing and Plagiarism Detection
2. Typologies
   - Paraphrasing Complexity
   - State of the Art
   - A Paraphrasing Typology
3. Corpora
   - State of the Art
   - CoCo Interface
4. NLP Approaches
   - State of the Art
5. The WRPA System
   - Presentation
   - Methodology
   - Evaluation and Results
   - Conclusions and Future Work

WRPA–Relational Paraphrase Aquisition from Wikipedia



Structured information

↓ ↓

{Source} {Target}

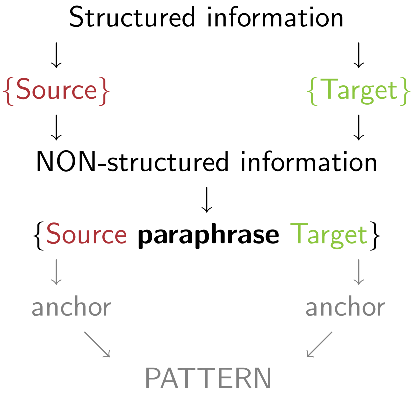
↓ ↓

NON-structured information

↓

{Source **paraphrase** Target}

WRPA–Relational Paraphrase Aquisition from Wikipedia



Structured information

{Source}

{Target}

NON-structured information

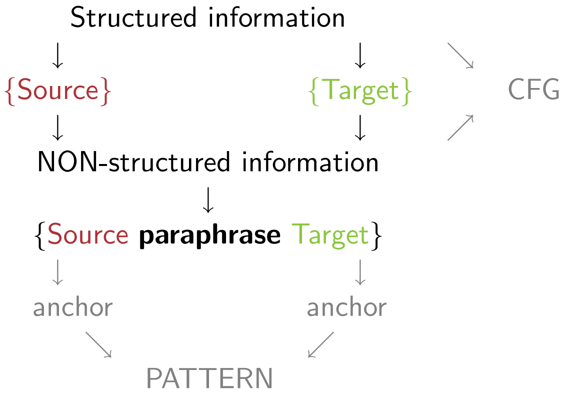
{Source **paraphrase** Target}

anchor

anchor

PATTERN

WRPA–Relational Paraphrase Aquisition from Wikipedia



Structured information

{Source}

{Target}

CFG

NON-structured information

{Source **paraphrase** Target}

anchor

anchor

PATTERN

WRPA–Relational Paraphrase Aquisition from Wikipedia

{**text**} [X] {**text**} Y {**text**} [Z] {**text**}

WRPA–Relational Paraphrase Aquisition from Wikipedia

{**text**}	[X]	{**text**}	Y	{**text**}	[Z]	{**text**}
↓	↓	↓	↓	↓	↓	↓
paraphr.	Source	**paraphr.**	Target	**paraphr.**	Compl. info	**paraphr.**

WRPA–Relational Paraphrase Aquisition from Wikipedia

{**text**}	[X]	{**text**}	Y	{**text**}	[Z]	{**text**}
↓	↓	↓	↓	↓	↓	↓
paraphr.	Source	**paraphr.**	Target	**paraphr.**	Compl. info	**paraphr.**
	Person		Date of birth Place of birth Date of death		–	

WRPA–Relational Paraphrase Aquisition from Wikipedia

{**text**}	[X]	{**text**}	Y	{**text**}	[Z]	{**text**}
↓	↓	↓	↓	↓	↓	↓
paraphr.	Source	**paraphr.**	Target	**paraphr.**	Compl. info	**paraphr.**
	Person		Date of birth Place of birth Date of death		–	
	Author		Work		Date	

WRPA–Relational Paraphrase Aquisition from Wikipedia

{**text**}	[X]	{**text**}	Y	{**text**}	[Z]	{**text**}
↓	↓	↓	↓	↓	↓	↓
paraphr.	Source	**paraphr.**	Target	**paraphr.**	Compl. info	**paraphr.**
	Person		Date of birth Place of birth Date of death		–	(English)
	Author		Work		Date	(Spanish)

WRPA–Relational Paraphrase Aquisition from Wikipedia

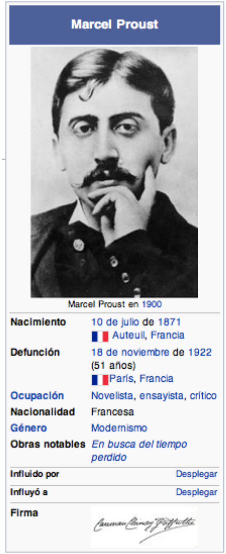
Authorship

- Complex relation (painters, sculptors, architects, writers, composers, singer-songwriters, directors, philosophers, inventors and scientists).
- Rich casuistic of manifestations.

  → Appropriateness of this relation for research in paraphrasing.

Wikipedia components and sections relevant to our work:

- Categories
- Infoboxes
- Work section
- Work page



Marcel Proust

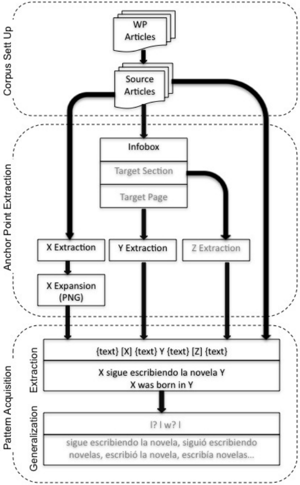
Marcel Proust en 1900

Nacimiento	10 de julio de 1871 Auteuil, Francia
Defunción	18 de noviembre de 1922 (51 años) París, Francia
Ocupación	Novelista, ensayista, crítico
Nacionalidad	Francesa
Género	Modernismo
Obras notables	*En busca del tiempo perdido*
Influido por	Desplegar
Influyó a	Desplegar
Firma	

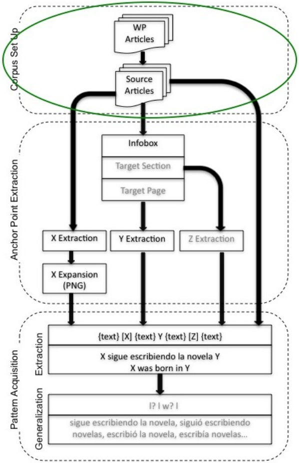
Wikipedia components and sections relevant to our work:

- Categories
- Infoboxes
- Work section
- Work page

System scheme



Corpus Set Up



Corpus Set Up

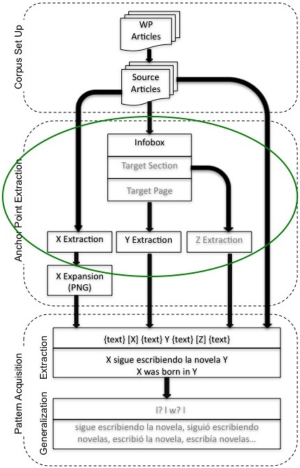
Date of birth
Place of birth → Persons
Date of death
Authorship → Authors

Corpus Set Up

Date of birth Place of birth Date of death	→	Persons
Authorship	→	Authors

*A category of authors does not exist in the Spanish Wikipedia → done manually

Anchor Point Extraction



Anchor Point Extraction

X Extraction

- Titles
- Redirection pages

Y and Z Extraction

Person:

- 'Date of birth', 'place of birth' and 'date of death' attributes in infoboxes.

Authorship:

- 'Work' attribute in infoboxes.
- Work section
- Work page

Anchor Point Extraction

X Extraction

- Titles
- Redirection pages

Y and Z Extraction

Person:

- 'Date of birth', 'place of birth' and 'date of death' attributes in infoboxes.

Authorship:

- 'Work' attribute in infoboxes.
- Work section
- Work page

*Attributes are expressed in different ways → done manually/KBP track (TAC 2010)

Anchor Point Extraction

X Extraction

- Titles
- Redirection pages

Y and Z Extraction

Person:

- 'Date of birth', 'place of birth' and 'date of death' attribute in infoboxes.

Authorship:

- 'Work' attribute in infoboxes.
- Work sections
- Work pages

∗ Attributes are univalued.

∗ The English Wikipedia is very extensive.

∗ There is a relatively high number of infoboxes in person pages (34 %).

∗ The corresponding attributes generally appear in the infoboxes.

∗ Attributes are multivalued.

∗ The Spanish Wikipedia is smaller.

∗ Most author pages lack an infobox.

∗ Most infoboxes lack a work attribute.

∗ Infoboxes only contain the most important works (2 approx.).

Anchor Point Extraction

X Extraction

- Titles
- Redirection pages

Y and Z Extraction

Person:

- 'Date of birth', 'place of birth' and 'date of death' attribute in infoboxes.

Authorship:

- 'Work' attribute in infoboxes.
- Work sections
- Work pages

* Attributes are univalued.
* The English Wikipedia is very extensive.
* There is a relatively high number of infoboxes in person pages (34 %).
* The corresponding attributes generally appear in the infoboxes.

* Attributes are multivalued.
* The Spanish Wikipedia is smaller.
* Most author pages lack an infobox.
* Most infoboxes lack a work attribute.
* Infoboxes only contain the most important works (2 approx.).

Anchor Point Extraction

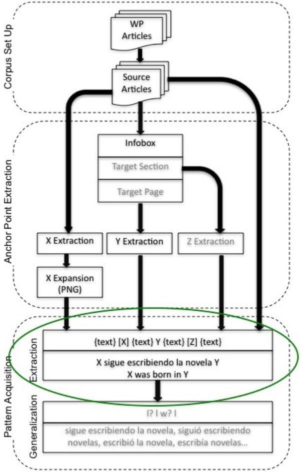
Paul Auster:

Infobox	0
Work section	33 (25 correct)
Work page	18 (all correct)

Anchor Point Extraction

<table>
<tr><th>X</th><th>Y(Z)</th></tr>
<tr><td colspan="2"><b>Authorship</b></td></tr>
<tr><td>Canaletto</td><td>La Riva degli Schiavoni (1730-31)</td></tr>
<tr><td>Edgar Allan Poe</td><td>[[Eureka]]</td></tr>
<tr><td>Luis Eduardo Aute</td><td>Templo de carne (1986)</td></tr>
<tr><td colspan="2"><b>Date of birth</b></td></tr>
<tr><td>David Kaye</td><td>[[October 14]], [[1964]]</td></tr>
<tr><td>Sara Rue</td><td>1979|01|26</td></tr>
<tr><td>Joan of Arc</td><td>c. [[1412]]</td></tr>
<tr><td colspan="2"><b>Place of birth</b></td></tr>
<tr><td>Giovanni Branca</td><td>[[San Angelo]] in [[Lizzola]], [[Pesaro]]</td></tr>
<tr><td>Grigore Preoteasa</td><td>[[Bucharest]], [[Romania]]</td></tr>
<tr><td>Tomas Plekanec</td><td>[[Kladno]], [[Czech Republic |CZE]]</td></tr>
</table>

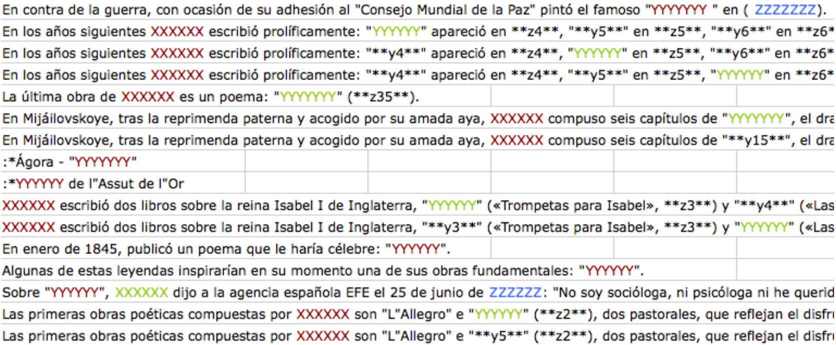
Pattern Extraction



Pattern Extraction

. {**text**} [X] {**text**} Y {**text**} [Z] {**text**} **.**

Pattern Extraction



En contra de la guerra, con ocasión de su adhesión al "Consejo Mundial de la Paz" pintó el famoso "YYYYYYY " en (ZZZZZZZ).

En los años siguientes XXXXXX escribió prolíficamente: "YYYYYY" apareció en **z4**, "**y5**" en **z5**, "**y6**" en **z6*

En los años siguientes XXXXXX escribió prolíficamente: "**y4**" apareció en **z4**, "YYYYYY" en **z5**, "**y6**" en **z6*

En los años siguientes XXXXXX escribió prolíficamente: "**y4**" apareció en **z4**, "**y5**" en **z5**, "YYYYYY" en **z6*

La última obra de XXXXXX es un poema: "YYYYYYY" (**z35**).

En Mijáilovskoye, tras la reprimenda paterna y acogido por su amada aya, XXXXXX compuso seis capítulos de "YYYYYYY", el dra

En Mijáilovskoye, tras la reprimenda paterna y acogido por su amada aya, XXXXXX compuso seis capítulos de "**y15**", el dra

:*Ágora - "YYYYYYY"

:*YYYYYY de l"Assut de l"Or

XXXXXX escribió dos libros sobre la reina Isabel I de Inglaterra, "YYYYYY" («Trompetas para Isabel», **z3**) y "**y4**" («Las

XXXXXX escribió dos libros sobre la reina Isabel I de Inglaterra, "**y3**" («Trompetas para Isabel», **z3**) y "YYYYYY" («Las

En enero de 1845, publicó un poema que le haría célebre: "YYYYYY".

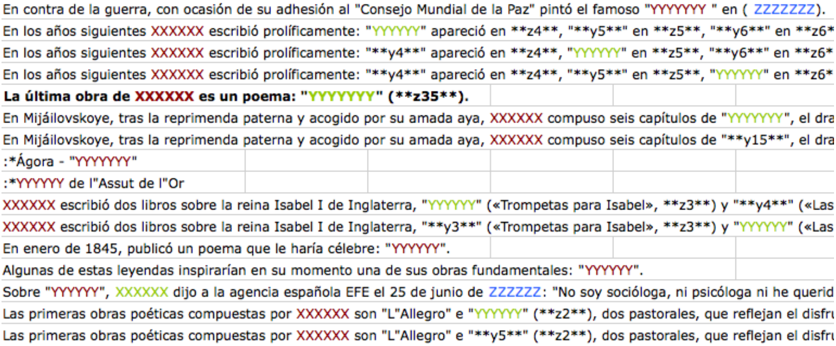
Algunas de estas leyendas inspirarían en su momento una de sus obras fundamentales: "YYYYYY".

Sobre "YYYYYY", XXXXXX dijo a la agencia española EFE el 25 de junio de ZZZZZZ: "No soy socióloga, ni psicóloga ni he querid

Las primeras obras poéticas compuestas por XXXXXX son "L"Allegro" e "YYYYYY" (**z2**), dos pastorales, que reflejan el disfr

Las primeras obras poéticas compuestas por XXXXXX son "L"Allegro" e "**y5**" (**z2**), dos pastorales, que reflejan el disfr

Pattern Extraction



En contra de la guerra, con ocasión de su adhesión al "Consejo Mundial de la Paz" pintó el famoso "YYYYYYY " en (ZZZZZZZ).

En los años siguientes XXXXXX escribió prolíficamente: "YYYYYY" apareció en **z4**, "**y5**" en **z5**, "**y6**" en **z6*

En los años siguientes XXXXXX escribió prolíficamente: "**y4**" apareció en **z4**, "YYYYYY" en **z5**, "**y6**" en **z6*

En los años siguientes XXXXXX escribió prolíficamente: "**y4**" apareció en **z4**, "**y5**" en **z5**, "YYYYYY" en **z6*

La última obra de XXXXXX es un poema: "YYYYYYY" (**z35**).

En Mijáilovskoye, tras la reprimenda paterna y acogido por su amada aya, XXXXXX compuso seis capítulos de "YYYYYYY", el dra

En Mijáilovskoye, tras la reprimenda paterna y acogido por su amada aya, XXXXXX compuso seis capítulos de "**y15**", el dra

:*Ágora - "YYYYYYY"

:*YYYYYY de l"Assut de l"Or

XXXXXX escribió dos libros sobre la reina Isabel I de Inglaterra, "YYYYYY" («Trompetas para Isabel», **z3**) y "**y4**" («Las

XXXXXX escribió dos libros sobre la reina Isabel I de Inglaterra, "**y3**" («Trompetas para Isabel», **z3**) y "YYYYYY" («Las

En enero de 1845, publicó un poema que le haría célebre: "YYYYYY".

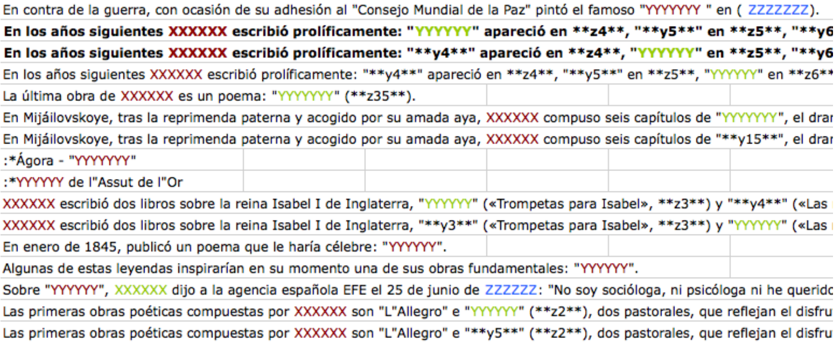
Algunas de estas leyendas inspirarían en su momento una de sus obras fundamentales: "YYYYYY".

Sobre "YYYYYY", XXXXXX dijo a la agencia española EFE el 25 de junio de ZZZZZZ: "No soy socióloga, ni psicóloga ni he querid

Las primeras obras poéticas compuestas por XXXXXX son "L"Allegro" e "YYYYYY" (**z2**), dos pastorales, que reflejan el disfru

Las primeras obras poéticas compuestas por XXXXXX son "L"Allegro" e "**y5**" (**z2**), dos pastorales, que reflejan el disfru

Pattern Extraction



En contra de la guerra, con ocasión de su adhesión al "Consejo Mundial de la Paz" pintó el famoso "YYYYYYY " en (ZZZZZZZ).

En los años siguientes XXXXXX escribió prolíficamente: "YYYYYY" apareció en **z4, "**y5**" en **z5**, "**y6**

En los años siguientes XXXXXX escribió prolíficamente: "y4**" apareció en **z4**, "YYYYYY" en **z5**, "**y6**

En los años siguientes XXXXXX escribió prolíficamente: "**y4**" apareció en **z4**, "**y5**" en **z5**, "YYYYYY" en **z6**

La última obra de XXXXXX es un poema: "YYYYYYY" (**z35**).

En Mijáilovskoye, tras la reprimenda paterna y acogido por su amada aya, XXXXXX compuso seis capítulos de "YYYYYYY", el drar

En Mijáilovskoye, tras la reprimenda paterna y acogido por su amada aya, XXXXXX compuso seis capítulos de "**y15**", el drar

:*Ágora - "YYYYYYY"

:*YYYYYY de l"Assut de l"Or

XXXXXX escribió dos libros sobre la reina Isabel I de Inglaterra, "YYYYYY" («Trompetas para Isabel», **z3**) y "**y4**" («Las

XXXXXX escribió dos libros sobre la reina Isabel I de Inglaterra, "**y3**" («Trompetas para Isabel», **z3**) y "YYYYYY" («Las

En enero de 1845, publicó un poema que le haría célebre: "YYYYYY".

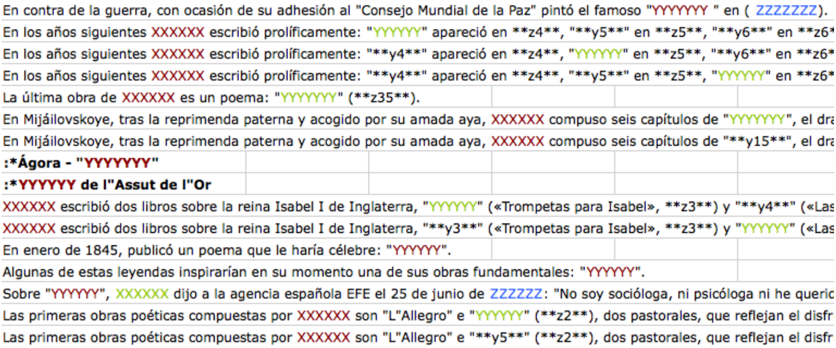
Algunas de estas leyendas inspirarían en su momento una de sus obras fundamentales: "YYYYYY".

Sobre "YYYYYY", XXXXXX dijo a la agencia española EFE el 25 de junio de ZZZZZZ: "No soy socióloga, ni psicóloga ni he querido

Las primeras obras poéticas compuestas por XXXXXX son "L"Allegro" e "YYYYYY" (**z2**), dos pastorales, que reflejan el disfru

Las primeras obras poéticas compuestas por XXXXXX son "L"Allegro" e "**y5**" (**z2**), dos pastorales, que reflejan el disfru

Pattern Extraction



En contra de la guerra, con ocasión de su adhesión al "Consejo Mundial de la Paz" pintó el famoso "YYYYYYY " en (ZZZZZZZ).

En los años siguientes XXXXXX escribió prolíficamente: "YYYYYY" apareció en **z4**, "**y5**" en **z5**, "**y6**" en **z6*

En los años siguientes XXXXXX escribió prolíficamente: "**y4**" apareció en **z4**, "YYYYYY" en **z5**, "**y6**" en **z6*

En los años siguientes XXXXXX escribió prolíficamente: "**y4**" apareció en **z4**, "**y5**" en **z5**, "YYYYYY" en **z6*

La última obra de XXXXXX es un poema: "YYYYYYY" (**z35**).

En Mijáilovskoye, tras la reprimenda paterna y acogido por su amada aya, XXXXXX compuso seis capítulos de "YYYYYYY", el dra

En Mijáilovskoye, tras la reprimenda paterna y acogido por su amada aya, XXXXXX compuso seis capítulos de "**y15**", el dra

:*Ágora - "YYYYYYY"

:*YYYYYY de l"Assut de l"Or

XXXXXX escribió dos libros sobre la reina Isabel I de Inglaterra, "YYYYYY" («Trompetas para Isabel», **z3**) y "**y4**" («Las

XXXXXX escribió dos libros sobre la reina Isabel I de Inglaterra, "**y3**" («Trompetas para Isabel», **z3**) y "YYYYYY" («Las

En enero de 1845, publicó un poema que le haría célebre: "YYYYYY".

Algunas de estas leyendas inspirarían en su momento una de sus obras fundamentales: "YYYYYY".

Sobre "YYYYYY", XXXXXX dijo a la agencia española EFE el 25 de junio de ZZZZZZ: "No soy socióloga, ni psicóloga ni he queric

Las primeras obras poéticas compuestas por XXXXXX son "L"Allegro" e "YYYYYY" (**z2**), dos pastorales, que reflejan el disfr

Las primeras obras poéticas compuestas por XXXXXX son "L"Allegro" e "**y5**" (**z2**), dos pastorales, que reflejan el disfr

Pattern Extraction

$$. \ \{\textbf{text}\} \ [\mathrm{X}] \ \{\textbf{text}\} \ \mathrm{Y} \ \{\textbf{text}\} \ [\mathrm{Z}] \ \{\textbf{text}\} \ .$$

$$\downarrow$$

$$. \ \mathrm{X} \ \{\textbf{text}\} \ \mathrm{Y} \ .$$

Pattern Extraction

Authorship

1 X sigue escribiendo la novela Y
2 X comenzó a grabar su álbum debut, “Y”
3 X dirigió ‘Educando a Rita”(YEAR) y “Y”

Date of birth

4 X known as PERSON was born in Y
5 X was born in Y
6 X was born in PLACE on Y

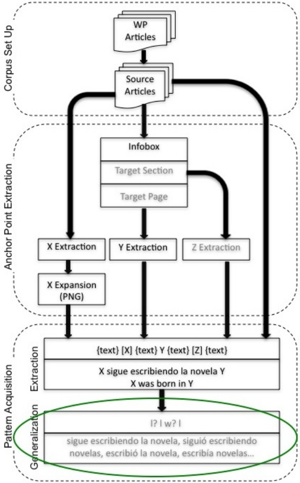
Date of death

7 X DATE-Y
8 X DATE to Y
9 X PERSON DATE - Y

Place of death

10 X born DATE in Y
11 X DATE in Y
12 X DATE Y

Pattern Generalitzation



Corpus Set Up

WP Articles

Source Articles

Anchor Point Extraction

Infobox

Target Section

Target Page

X Extraction

Y Extraction

Z Extraction

X Expansion (PNG)

Pattern Acquisition

Extraction

{text} [X] {text} Y {text} [Z] {text}

X sigue escribiendo la novela Y
X was born in Y

Generalization

l? l w? l

sigue escribiendo la novela, siguió escribiendo novelas, escribió la novela, escribía novelas...

Pattern Generalitzation

To gather the variants of a generic pattern (only for authorship).

- To lemmatize and PoS tag the patterns (Freeling).
- To represent each pattern as a sequence of <word, lemma, PoS> tuples.

Word	sigue	escribiendo	la	novela
Lemma	seguir	escribir	el	novela
PoS	v	v	da	nc

- To use an A* approach until achieving *n* matches with other patterns.

Pattern Generalitzation

- Each state is represented as a sequence of tuples consisting of a token (word, lemma or PoS) and a condition (obligatory, skippable or omitted):

Initial state	<sigue:w> <escribiendo:w> <la:w> <novela:w>
Matching pattern	sigue escribiendo la novela
Generalization	<sigue:l?> <escribiendo:l> <la:w?> <novela:l>
Matching patterns	1) sigue escribiendo la novela 2) siguió escribiendo novelas 3) escribió la novela 4) escribía novelas

Outline

1. Paraphrasing and Plagiarism Detection
2. Typologies
   - Paraphrasing Complexity
   - State of the Art
   - A Paraphrasing Typology
3. Corpora
   - State of the Art
   - CoCo Interface
4. NLP Approaches
   - State of the Art
5. The WRPA System
   - Presentation
   - Methodology
   - Evaluation and Results
   - Conclusions and Future Work

Evaluation

Precision

To apply each pattern to its original page and verify whether the output is a Y (or a variant) of the corresponding infobox (correct Ys).

$$\frac{Num.\ of\ correct\ Ys}{Num.\ of\ extracted\ Ys}$$

Recall

Person: To apply the pattern to its original page and verify whether the output is a Y (it can be applied to that page).
Conservative assumption: all the pages contain a Y.

$$\frac{Num.\ of\ extracted\ Ys}{Num.\ of\ pages\ of\ the\ corpus}$$

Baseline

Person: The most frequent pattern.

Results

			Precision	Recall	F1
Person	**Date of birth**	X born Y	0.95	0.57	0.71
		X Y	0.80	0.12	0.21
		Top 8 patterns	0.92	0.75	0.83
		Baseline	0.95	0.57	0.71
	Date of death	X DATE-Y	0.95	0.21	0.34
		X PERSON DATE-Y	0.96	0.10	0.18
		Top 3 patterns	0.95	0.42	0.58
		Baseline	0.95	0.21	0.34
	Place of birth	X born DATE in Y	0.98	0.13	0.23
		X DATE in Y	0.94	0.10	0.18
		Top 3 patterns	0.92	0.26	0.41
		Baseline	0.98	0.13	0.23
Authorship	<pintó:w><su:w?><cuadro:w?>		0.46	–	–
	<pintó:w><su:w?><primera:l?><obra:w?>		0.49	–	–

Results

			Precision	**Recall**	**F1**
Person	**Date of birth**	X born Y	0.95	0.57	0.71
		X Y	0.80	0.12	0.21
		Top 8 patterns	0.92	0.75	0.83
		Baseline	0.95	0.57	0.71
	Date of death	X DATE-Y	0.95	0.21	0.34
		X PERSON DATE-Y	0.96	0.10	0.18
		Top 3 patterns	0.95	0.42	0.58
		Baseline	0.95	0.21	0.34
	Place of birth	X born DATE in Y	0.98	0.13	0.23
		X DATE in Y	0.94	0.10	0.18
		Top 3 patterns	0.92	0.26	0.41
		Baseline	0.98	0.13	0.23
Authorship	<pintó:w><su:w?><cuadro:w?>		0.46	–	–
	<pintó:w><su:w?><primera:l?><obra:w?>		0.49	–	–

Results

			Precision	Recall	F1
Person	**Date of birth**	X born Y	0.95	0.57	0.71
		X Y	0.80	0.12	0.21
		Top 8 patterns	0.92	0.75	0.83
		Baseline	0.95	0.57	0.71
	Date of death	X DATE-Y	0.95	0.21	0.34
		X PERSON DATE-Y	0.96	0.10	0.18
		Top 3 patterns	0.95	0.42	0.58
		Baseline	0.95	0.21	0.34
	Place of birth	X born DATE in Y	0.98	0.13	0.23
		X DATE in Y	0.94	0.10	0.18
		Top 3 patterns	0.92	0.26	0.41
		Baseline	0.98	0.13	0.23
Authorship	<pintó:w><su:w?><cuadro:w?>		0.46	–	–
	<pintó:w><su:w?><primera:l?><obra:w?>		0.49	–	–

Outline

1. Paraphrasing and Plagiarism Detection
2. Typologies
   - Paraphrasing Complexity
   - State of the Art
   - A Paraphrasing Typology
3. Corpora
   - State of the Art
   - CoCo Interface
4. NLP Approaches
   - State of the Art
5. The WRPA System
   - Presentation
   - Methodology
   - Evaluation and Results
   - Conclusions and Future Work

Conclusions

Finding good and numerous anchor points is essential in systems based on the Distributional Hypothesis.

Conclusions

Finding good and numerous anchor points is essential in systems based on the Distributional Hypothesis.

WRPA...

- ...guarantees the **quality** of these anchor points as they are directly extracted from structured and semantically labelled data.

Conclusions

Finding good and numerous anchor points is essential in systems based on the Distributional Hypothesis.

WRPA...

- ...guarantees the **quality** of these anchor points as they are directly extracted from structured and semantically labelled data.
- ...guarantees a large **quantity** of anchor points as they are not only extracted from infoboxes but also from section and work pages.

Conclusions

Finding good and numerous anchor points is essential in systems based on the Distributional Hypothesis.

WRPA...

- ...guarantees the **quality** of these anchor points as they are directly extracted from structured and semantically labelled data.
- ...guarantees a large **quantity** of anchor points as they are not only extracted from infoboxes but also from section and work pages.
- ...can deal with **complex relationships**.

Conclusions

Finding good and numerous anchor points is essential in systems based on the Distributional Hypothesis.

WRPA...

- ...guarantees the **quality** of these anchor points as they are directly extracted from structured and semantically labelled data.
- ...guarantees a large **quantity** of anchor points as they are not only extracted from infoboxes but also from section and work pages.
- ...can deal with **complex relationships**.
- ...is **language independent** and also independent of the **relation**.

Conclusions

Finding good and numerous anchor points is essential in systems based on the Distributional Hypothesis.

WRPA...

- ...guarantees the **quality** of these anchor points as they are directly extracted from structured and semantically labelled data.
- ...guarantees a large **quantity** of anchor points as they are not only extracted from infoboxes but also from section and work pages.
- ...can deal with **complex relationships**.
- ...is **language independent** and also independent of the **relation**.
- ...only relies on **Wikipedia** and **shallow linguistic processing**.

Future Work

- Recall and baseline for authorship.

Future Work

- Recall and baseline for authorship.
- Manual validation of paraphrase candidates: CoCo (España et al., 2009)
    - Do they express an authorship relation?
    - Establishement of paraphrasing relations between them in the framework of a typology.

Outline

1. Paraphrasing and Plagiarism Detection
2. Typologies
   - Paraphrasing Complexity
   - State of the Art
   - A Paraphrasing Typology
3. Corpora
   - State of the Art
   - CoCo Interface
4. NLP Approaches
   - State of the Art
5. The WRPA System
   - Presentation
   - Methodology
   - Evaluation and Results
   - Conclusions and Future Work

Thank you!

Questions?

marta.vila@ub.edu