

# Cross-Language Plagiarism Detection using a Multilingual Semantic Network

Marc Franco Salvador

<http://users.dsic.upv.es/~mfranco/>

Advisor: Paolo Rosso

Universitat Politècnica de Valencia

November 18, 2013



UNIVERSITAT  
POLITÈCNICA  
DE VALÈNCIA

Language Langue Linguaggio  
Языки Languages BRACHE  
**NLEL**  
Natural Language Engineering Lab  
Lingua LENGUAIGE

 MÀSTER EN INTEL·LENCIA ARTIFICIAL,  
RECONOCIMIENTO DE FORMAS E IMAGEN DIGITAL **MIARFID**



# Outline

- Introduction
- Related Work
- Knowledge Graphs
- Cross-Language Knowledge Graph Analysis
- Evaluation
- Conclusions and future work



# Outline

- **Introduction**
- Related Work
- Knowledge Graphs
- Cross-Language Knowledge Graph Analysis
- Evaluation
- Conclusions and future work



# Plagiarism

- Unauthorized use of the original content from authors.



# Plagiarism

- Unauthorized use of the original content from authors.
- The fact of presenting others' work or ideas as your own.



# Plagiarism

- Unauthorized use of the original content from authors.
- The fact of presenting others' work or ideas as your own.
- The deliberate use of someone else's original material without acknowledging its source.



# Plagiarism example

- “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.”
- “The strike began on May 29, and on June 1 the manufacturers met publicly to plan their response. They had two strategies.”



# Plagiarism example

- “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.”
- “The strike began on May 29, and on June 1 the manufacturers met publicly to plan their response. They had two strategies.”





# Cross-language plagiarism

- When the source of the plagiarism comes from another language.

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

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



# Cross-language plagiarism

- When the source of the plagiarism comes from another language.
- Copy and translate original content without acknowledging its source.

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

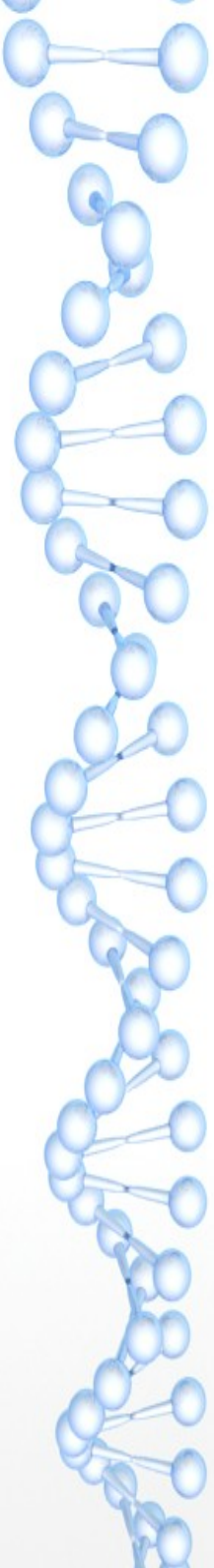
ES: *“La huelga comenzó oficialmente el 29 de mayo, y el 1 de junio los fabricantes se reunieron públicamente para planificar su resistencia.”*



# Cross-language plagiarism detection

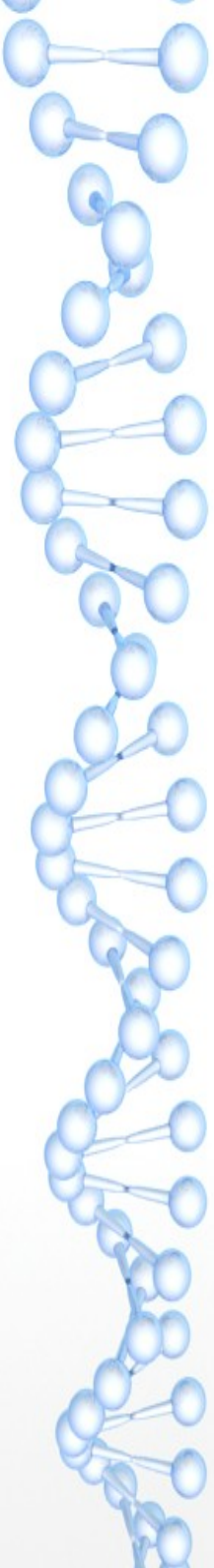
- It is the task which tries to find automatically the sections of text involved in plagiarism among documents in different languages.

# Motivation



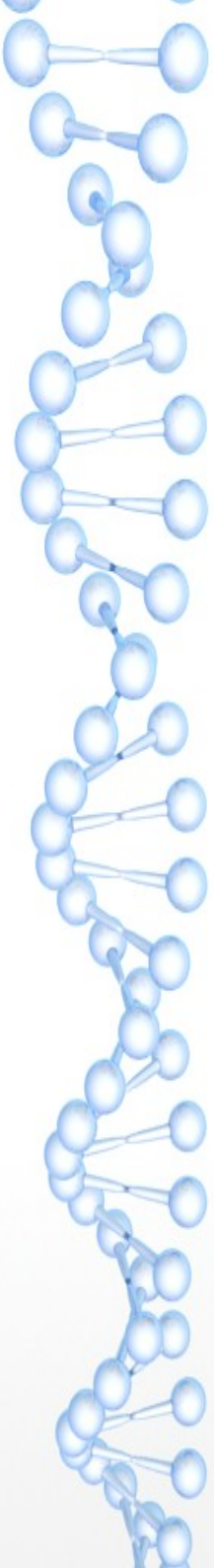
# Motivation

- Internet



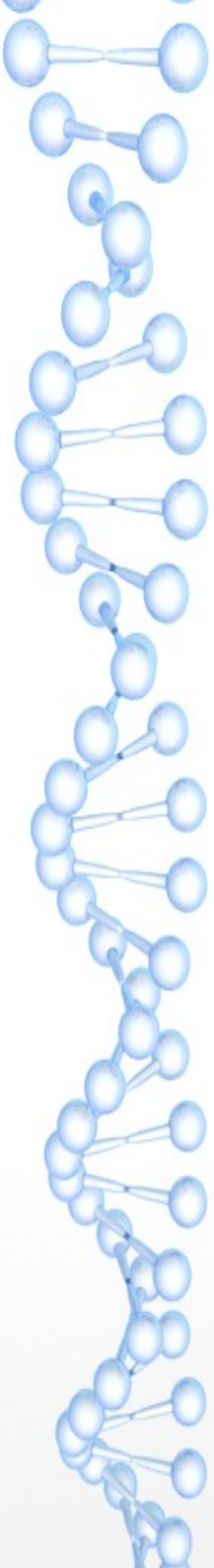
# Motivation

- Internet
- **Students**

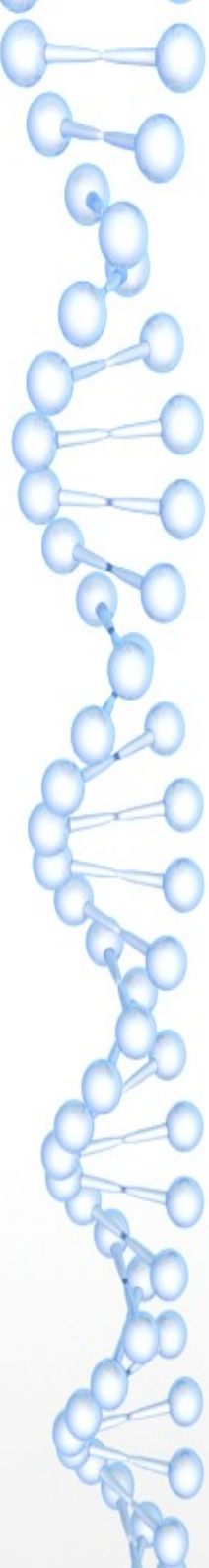


# Motivation

- Internet
- Students
- Literature



# Motivation

- 
- Internet
  - Students
  - Literature
  - Science

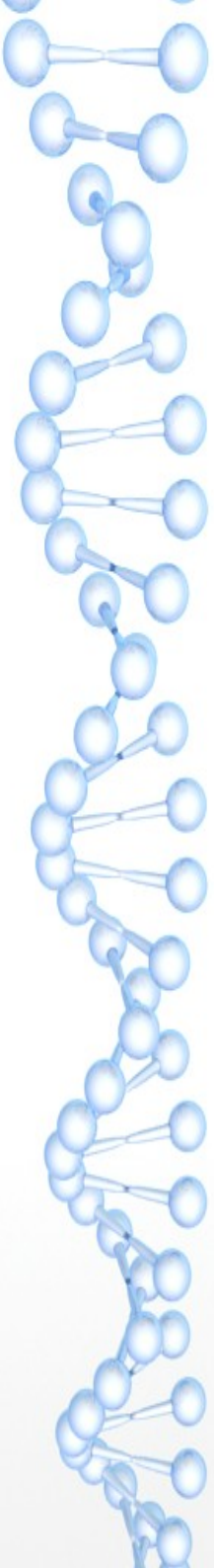




# Outline

- Introduction
- **Related Work**
- Knowledge Graphs
- Cross-Language Knowledge Graph Analysis
- Evaluation
- Conclusions and future work

# Cross-language retrieval models





# Cross-language retrieval models

- Models based on language syntax:
  - CL-CNG



# Cross-language retrieval models

- Models based on language syntax:
  - CL-CNG
- Models based on dictionaries, gazetteers, rules and thesauri:
  - CL-CTS, CL-VSM



# Cross-language retrieval models

- Models based on language syntax:
  - CL-CNG
- Models based on dictionaries, gazetteers, rules and thesauri:
  - CL-CTS, CL-VSM
- **Models based on comparable corpora:**
  - CL-ESA



# Cross-language retrieval models

- Models based on language syntax:
  - CL-CNG
- Models based on dictionaries, gazetteers, rules and thesauri:
  - CL-CTS, CL-VSM
- Models based on comparable corpora:
  - CL-ESA
- **Models based on parallel corpora:**
  - CL-ASA, CL-KCCA, CL-LSI



# Cross-language retrieval models

Potthast et al., 2011a, Gupta et al., 2012 and Barrón-Cedeño et al., 2013, have compared these models. **CL-ASA** and **CL-CNG** achieved the best performance.



# Cross-Language Character N-Grams

CL-CNG [McNamee and Mayfield, 2004] model achieves a remarkable performance in keyword retrieval for languages with lexical similarities.

Similarity between two documents  $d$  and  $d'$  is computed as follows:

$$S(d, d') = \frac{\vec{d} \cdot \vec{d}'}{\|\vec{d}\| \cdot \|\vec{d}'\|}$$





# Cross-Language Character N-Grams

- Model limitations:



# Cross-Language Character N-Grams

- Model limitations:
  - Languages must have lexical similarities.



# Cross-Language Character N-Grams

- Model limitations:
  - Languages must have lexical similarities:

IT: *“questa è una frase di esempio”*

ES: *“Esta es una frase de ejemplo”*



# Cross-Language Character N-Grams

- Model limitations:

- Languages must have lexical similarities:

IT: *“questa è una frase di esempio”*

{que, ues, est, sta, ... una, naf, afr, ... emp, mpi, pio}

ES: *“Esta es una frase de ejemplo”*

{est, sta, tae, ... una, naf, afr, ... emp, mpl, plo}



# Cross-Language Character N-Grams

- Model limitations:

- Languages must have lexical similarities:

IT: *“questa è una frase di esempio”*

{que, ues, est, sta, ... una, naf, afr, ... emp, mpi, pio}

ES: *“Esta es una frase de ejemplo”*

{est, sta, tae, ... una, naf, afr, ... emp, mpl, plo}



# Cross-Language Character N-Grams

- Model limitations:
  - Languages must have lexical similarities:

IT: *“questa è una frase di esempio”*

DE: *“dies ist ein beispielsatz”*



# Cross-Language Character N-Grams

- Model limitations:

- Languages must have lexical similarities:

IT: *“questa è una frase di esempio”*

{que, ues, est, sta, ... una, naf, afr, ... emp, mpi, pio}

DE: *“dies ist ein beispielsatz”*

{die, ies, esi, ... , sat, atz}



# Cross-Language Character N-Grams

- Model limitations:
  - Languages must have lexical similarities:

IT: *“questa è una frase di esempio”*

{que, ues, est, sta, ... una, naf, afr, ... emp, mpi, pio}

DE: *“dies ist ein beispielsatz”*

{die, ies, esi, ... , sat, atz}





# Cross-Language Alignment based Similarity Analysis

CL-ASA [Barrón-Cedeño et al., 2008] combines probabilistic translation, using a statistical bilingual dictionary and similarity analysis, aligning documents at word level.

Similarity between two documents  $d$  and  $d'$  is computed as follows:

$$S(d, d') = l(d, d') t(d|d')$$

*Length model:*

$$l(d, d') = \exp\left(-0.5\left(\frac{|d'|/|d| - \mu}{\sigma}\right)^2\right)$$

*Translation model:*

$$t(d|d') = \sum_{x \in d} \sum_{y \in d'} p(x, y)$$



# Cross-Language Alignment based Similarity Analysis

- Model limitations:



# Cross-Language Alignment based Similarity Analysis

- Model limitations:
  - Translated plagiarism cases must be exact copies of the original source.



# Cross-Language Alignment based Similarity Analysis

- Model limitations:
  - Translated plagiarism cases must be exact copies of the original source.

IT: *“questa è una frase di esempio”*

DE: *“dies ist ein beispielsatz”*



# Cross-Language Alignment based Similarity Analysis

- Model limitations:
  - Translated plagiarism cases must be exact copies of the original source.

IT: *“questa è una frase di esempio”*

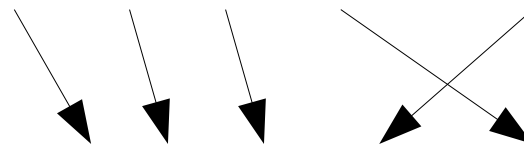
DE: *“dies ist ein beispielsatz”*

questa	dies	0.7	frase	satz	0.7
questa	dieses	0.1	...		
questa	das	0.2	...		
e	ist	0.8	esempio	beispiel	0.8
una	ein	0.9	esempio	muster	0.2

# Cross-Language Alignment based Similarity Analysis

- Model limitations:
  - Translated plagiarism cases must be exact copies of the original source.

IT: *“questa è una frase di esempio”*



DE: *“dies ist ein beispielsatz”*

questa	dies	0.7	frase	satz	0.7
questa	dieses	0.1	...		
questa	das	0.2	...		
e	ist	0.8	esempio	beispiel	0.8
una	ein	0.9	esempio	muster	0.2

Perfect alignment!



# Cross-Language Alignment based Similarity Analysis

- Model limitations:
  - Translated plagiarism cases must be exact copies of the original source.

IT: *“questa è una frase di esempio”*

EN: *“this is a demo text fragment”*



# Cross-Language Alignment based Similarity Analysis

- Model limitations:
  - Translated plagiarism cases must be exact copies of the original source.

IT: *“questa è una frase di esempio”*

EN: *“this is a demo text fragment”*

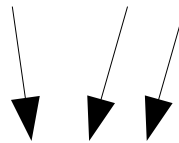
questa	this	0.7	frase	sentence	0.7
questa	that	0.3	frase	phrase	0.3
e	is	1.0	esempio	sample	0.6
...			esempio	example	0.4



# Cross-Language Alignment based Similarity Analysis

- Model limitations:
  - Translated plagiarism cases must be exact copies of the original source.

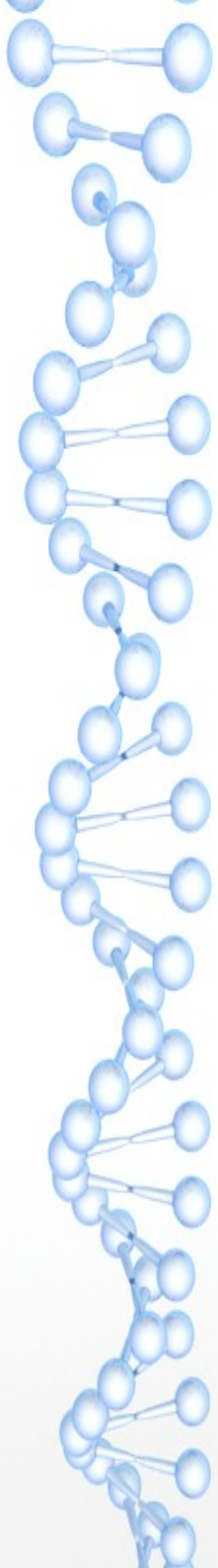
IT: “*questa è una frase di esempio*”



EN: “*this is a demo text fragment*”

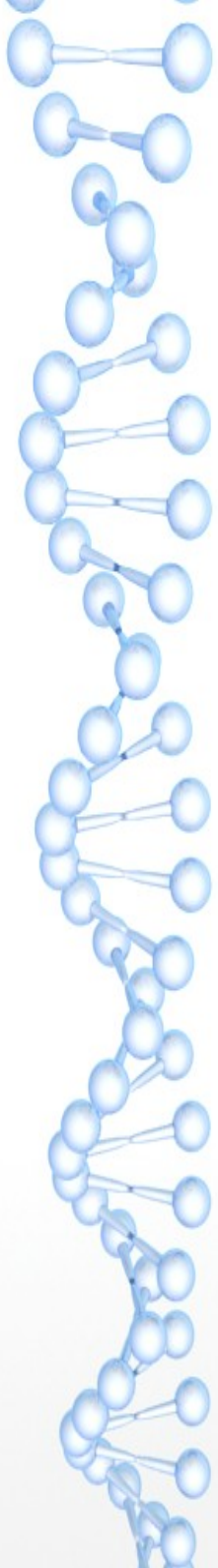
questa	this	0.7	frase	sentence	0.7
questa	that	0.3	frase	phrase	0.3
e	is	1.0	esempio	sample	0.6
...			esempio	example	0.4

we have lost the whole context!



The classical approaches have strong limitations to face the most common types of cross-language plagiarism.

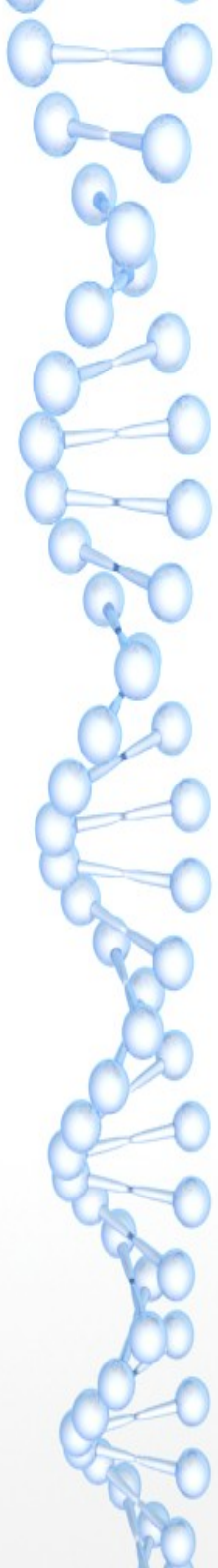
We need to go forward...



The classical approaches have strong limitations to face the most common types of cross-language plagiarism.

We need to go forward...

...to a semantic level.



Is there a language-independent way to model the context of a text fragment

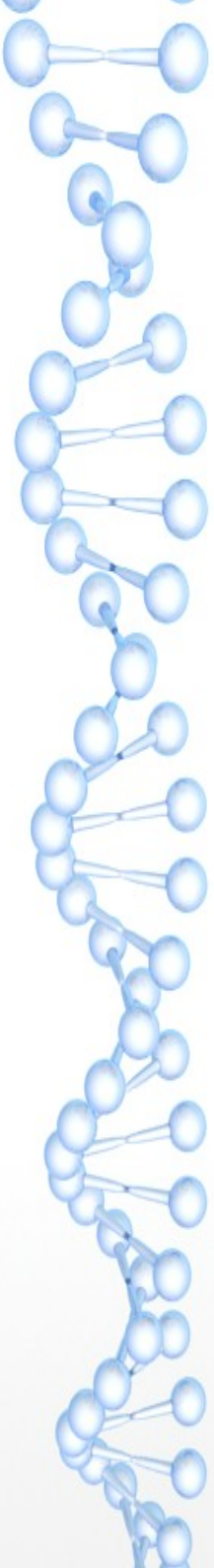




# Outline

- Introduction
- Related Work
- **Knowledge Graphs**
- Cross-Language Knowledge Graph Analysis
- Evaluation
- Conclusions and future work

# Knowledge graphs





# Knowledge graphs

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



# Knowledge graphs

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

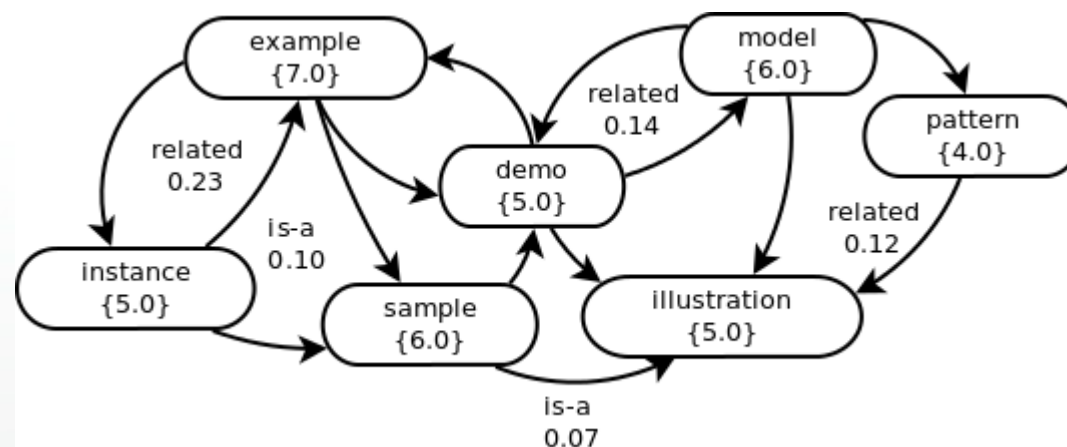
Concepts: {example, model}



# Knowledge graphs

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

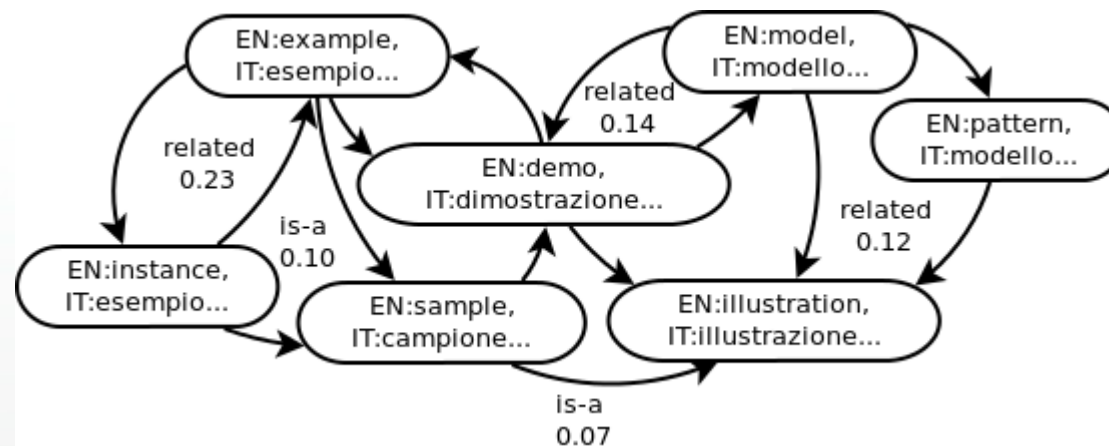
Concepts: {example, model}



# Knowledge graphs

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}





# BabelNet<sup>2.0</sup>

A very large multilingual encyclopedic dictionary and ontology

- Knowledge graphs are built using the multilingual semantic network BabelNet [Navigli and Ponzetto, 2012].



# BabelNet<sup>2.0</sup>

A very large multilingual **encyclopedia** and **ontology**

- Knowledge graphs are built using the multilingual semantic network BabelNet [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<sup>2.0</sup>

A very large multilingual **encyclopedic dictionary** and **ontology**

- Knowledge graphs are built using the multilingual semantic network BabelNet [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 2.0 covers **50 languages**.



# BabelNet<sup>2.0</sup>

A very large multilingual **encyclopedic dictionary** and **ontology**

- Knowledge graphs are built using the multilingual semantic network BabelNet [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 2.0 covers **50 languages**.
  - It integrates:
    - WordNet
    - Open Multilingual WordNet
    - Wikipedia
    - OmegaWiki





A very large multilingual **encyclopedic dictionary** and **ontology**

Creating a knowledge graph from a text fragment  $d$ :



A very large multilingual **encyclopedic dictionary** and **ontology**

Creating a knowledge graph from a text fragment  $d$ :

1<sup>st</sup> Lemmatize and POS tag the words in  $d$





A very large multilingual **encyclopedic dictionary** and **ontology**

Creating a knowledge graph from a text fragment  $d$ :

1<sup>st</sup> Lemmatize and POS tag the words in  $d$

2<sup>nd</sup> Get the synset list  $s$  containing that words



A very large multilingual **encyclopedic dictionary** and **ontology**

Creating a knowledge graph from a text fragment  $d$ :

1<sup>st</sup> Lemmatize and POS tag the words in  $d$

2<sup>nd</sup> Get the synset list  $s$  containing that words

3<sup>rd</sup> Search the paths between all the pairs of synsets in  $s$



A very large multilingual **encyclopedic dictionary** and **ontology**

Creating a knowledge graph from a text fragment  $d$ :

1<sup>st</sup> Lemmatize and POS tag the words in  $d$

2<sup>nd</sup> Get the synset list  $s$  containing that words

3<sup>rd</sup> Search the paths between all the pairs of synsets in  $s$

4<sup>th</sup> Merge the paths to obtain the graph  $G$



A very large multilingual **encyclopedic dictionary** and **ontology**

Creating a knowledge graph from a text fragment  $d$ :

1<sup>st</sup> Lemmatize and POS tag the words in  $d$

2<sup>nd</sup> Get the synset list  $s$  containing that words

3<sup>rd</sup> Search the paths between all the pairs of synsets in  $s$

4<sup>th</sup> Merge the paths to obtain the graph  $G$

5<sup>th</sup> Weight the concepts and relations of  $G$



# Outline

- Introduction
- Related Work
- Knowledge Graphs
- **Cross-Language Knowledge Graph Analysis**
- Evaluation
- Conclusions and future work



# Cross-Language Knowledge Graphs Analysis (CL-KGA)

- CL-KGA [Franco et al., 2013] provides a context model by generating knowledge graphs from suspicious and source words from documents.
- The similarity between two graphs  $G$  and  $G'$  is measured in a semantic graph space.

$$S(G, G') = S_c(G, G') + b S_r(G, G')$$

$$S_c(G, G') = \frac{2 \sum_{c \in G \cap G'} w(c)}{\sum_{c \in G} w(c) + \sum_{c \in G'} w(c)}$$

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



# Cross-Language Knowledge Graphs Analysis (CL-KGA)

IT: *“questa è una frase di esempio”*

EN: *“this is a demo text fragment”*





# Cross-Language Knowledge Graphs Analysis (CL-KGA)

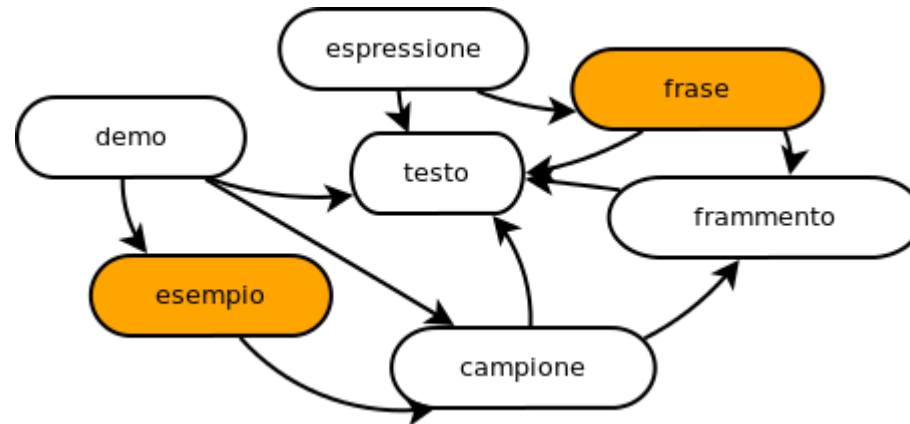
IT: “*questa è una frase di esempio*”

EN: “*this is a demo text fragment*”

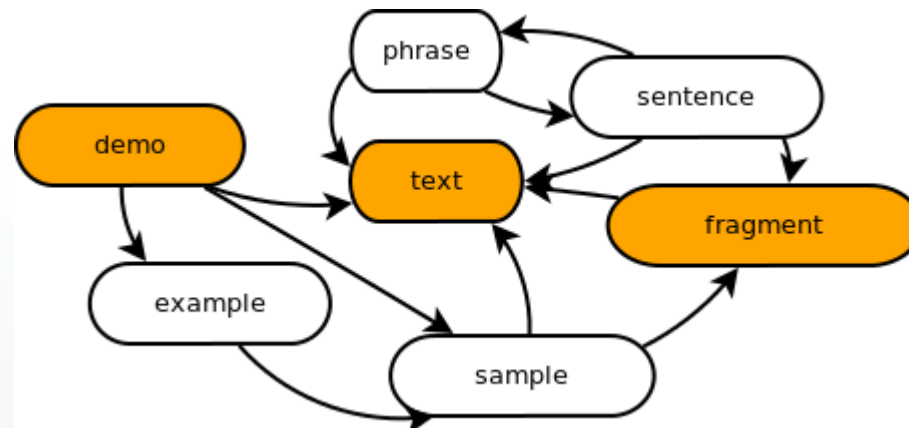


# Cross-Language Knowledge Graphs Analysis (CL-KGA)

IT: “questa è una *frase di esempio*”

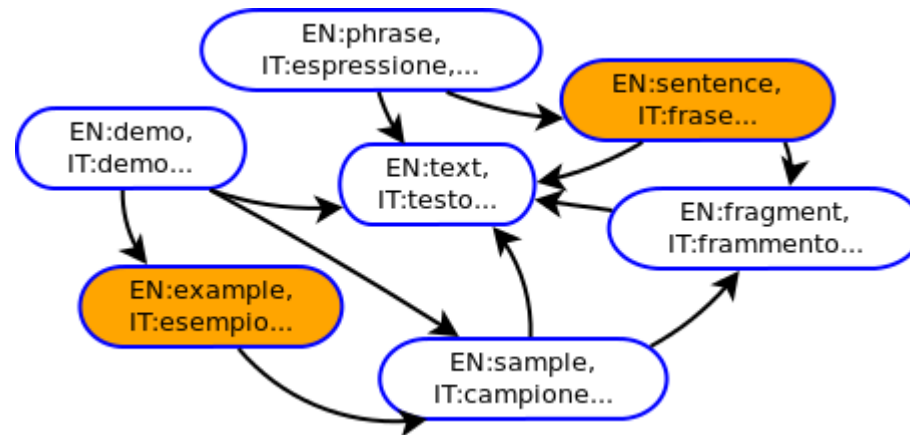


EN: “this is a *demo text fragment*”

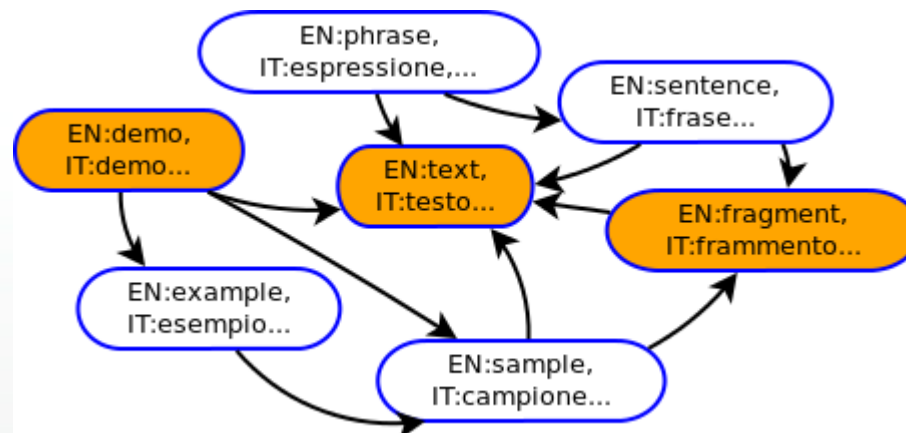


# Cross-Language Knowledge Graphs Analysis (CL-KGA)

IT: “*questa è una frase di esempio*”



EN: “*this is a demo text fragment*”





# Outline

- Introduction
- Related Work
- Knowledge Graphs
- Cross-Language Knowledge Graph Analysis
- **Evaluation**
- Conclusions and future work

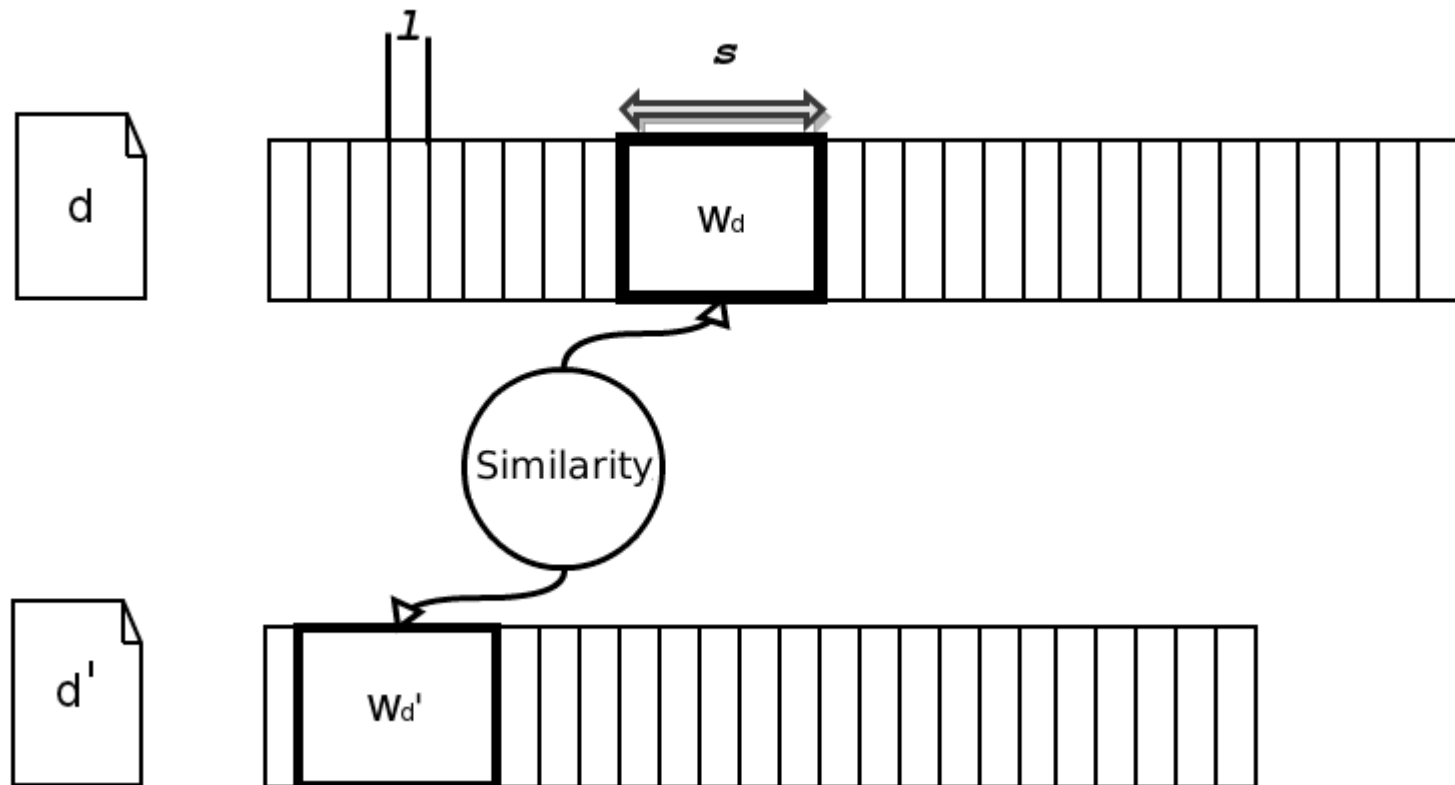


# Evaluation

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 of document** in  $D'$  from  $D$ .

# 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 of document** in  $D'$  from  $D$ .





# Evaluation

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



# Evaluation

Corpus: 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			5142
Automatic translation + Manual correction			433



# Evaluation

Automatic plagiarism case:

- 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.”
- ES: “La huelga comenzó oficialmente el 29 de mayo, y el 1 de junio los fabricantes se reunieron públicamente para planificar su resistencia. Sus estrategias se llevaron a cabo en dos frentes”





# Evaluation

Automatic plagiarism case + manual correction:

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

# Evaluation

Models:

- CL-C3G
- CL-ASA<sub>IBM M1</sub>
- CL-ASA<sub>BN</sub>
- CL-KGA



# Evaluation

Measures:

- **Recall** (at character level)
- **Precision** (at character level)



# Evaluation

Measures:

- **Recall** (at character level)
- **Precision** (at character level)
- **Granularity**: measures the error when detectors report overlapping or multiple detections for a single plagiarism case. The best possible value is 1.



# Evaluation

Measures:

- **Recall** (at character level)
- **Precision** (at character level)
- **Granularity**: measures the error when detectors report overlapping or multiple detections for a single plagiarism case. The best possible value is 1.

*d* : “questa è una frase di esempio”

*d'* : “this is a demo text fragment”



# Evaluation

Measures:

- **Recall** (at character level)
- **Precision** (at character level)
- **Granularity**: measures the error when detectors report overlapping or multiple detections for a single plagiarism case. The best possible value is 1.

*d* : “questa è una frase di esempio”

*d'* : “this is a demo text fragment”

Found 1 plagiarism case:

“questa è una frase di esempio” = “this is a demo text fragment”



# Evaluation

Measures:

- **Recall** (at character level)
- **Precision** (at character level)
- **Granularity**: measures the error when detectors report overlapping or multiple detections for a single plagiarism case. The best possible value is 1.

*d* : “questa è una frase di esempio”

*d'* : “this is a demo text fragment”

Found 1 plagiarism case:

“questa è una frase di esempio” = “this is a demo text fragment”

Granularity = 1



# Evaluation

Measures:

- **Recall** (at character level)
- **Precision** (at character level)
- **Granularity**: measures the error when detectors report overlapping or multiple detections for a single plagiarism case. The best possible value is 1.

*d* : “questa è una frase di esempio”

*d'* : “this is a demo text fragment”

**Found 2 plagiarism cases:**

- “questa è” = “this is”
- “frase di esempio” = “demo text fragment”





# Evaluation

Measures:

- **Recall** (at character level)
- **Precision** (at character level)
- **Granularity**: measures the error when detectors report overlapping or multiple detections for a single plagiarism case. The best possible value is 1.

*d* : “questa è una frase di esempio”

*d'* : “this is a demo text fragment”

**Found 2 plagiarism cases:**

- “questa è” = “this is”
- “frase di esempio” = “demo text fragment”

**Granularity** ↑↑



# Evaluation

Measures:

- **Recall** (at character level)
- **Precision** (at character level)
- **Granularity**: measures the error when detectors report overlapping or multiple detections for a single plagiarism case.
- **Plagdet**: is a combination of the previous measures to obtain an **overall score for plagiarism detection**:

$$plagdet(S, R) = \frac{F_1}{\log_2(1 + granularity(S, R))}$$

where  $S$  is the set of plagiarism cases in the corpus and  $R$  is the set of plagiarism cases reported by the detector.



# Evaluation

DE-EN results:

Model	Plagdet	Recall	Precision	Granularity
<b>CL-KGA</b>	<b>0.514</b>	<b>0.443</b>	<b>0.631</b>	<b>1.018</b>
CL-ASA <sub>IBM M1</sub>	0.406	0.344	0.604	1.113
CL-ASA <sub>BN</sub>	0.289	0.222	0.595	1.172
CL-C3G	0.078	0.047	0.330	1.089



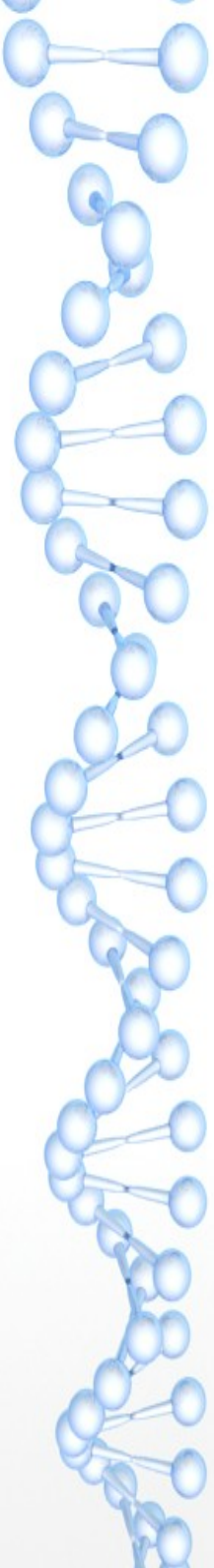
# Evaluation

ES-EN results:

Model	Plagdet	Recall	Precision	Granularity
<b>CL-KGA</b>	<b>0.599</b>	<b>0.525</b>	<b>0.703</b>	<b>1.004</b>
CL-ASA <sub>BN</sub>	0.554	0.491	0.663	1.015
CL-ASA <sub>IBM M1</sub>	0.517	0.448	0.689	1.071
CL-C3G	0.170	0.128	0.617	1.372

# Evaluation

Differences detecting automatic VS manual translations:



# Evaluation

Differences detecting automatic VS manual translations:

Model	DE-EN				ES-EN			
	Recall		Precision		Recall		Precision	
	automatic	manual	automatic	manual	automatic	manual	automatic	manual
<b>CL-KGA</b>	<b>.538</b>	<b>.247</b>	<b>.698</b>	<b>.098</b>	<b>.601</b>	<b>.221</b>	<b>.774</b>	<b>.098</b>
CL-ASA <sub>IBM M1</sub>	.538	.126	.642	.041	.596	.180	.741	.068
CL-ASA <sub>BN</sub>	.472	.092	.631	.033	.599	.198	.720	.076

# Evaluation

Differences detecting automatic VS manual translations:

Model	DE-EN				ES-EN			
	Recall		Precision		Recall		Precision	
	automatic	manual	automatic	manual	automatic	manual	automatic	manual
<b>CL-KGA</b>	<b>.538</b>	<b>.247</b>	<b>.698</b>	<b>.098</b>	<b>.601</b>	<b>.221</b>	<b>.774</b>	<b>.098</b>
CL-ASA <sub>IBM M1</sub>	.538	.126	.642	.041	.596	.180	.741	.068
CL-ASA <sub>BN</sub>	.472	.092	.631	.033	.599	.198	.720	.076

number of manual cases ↓↓



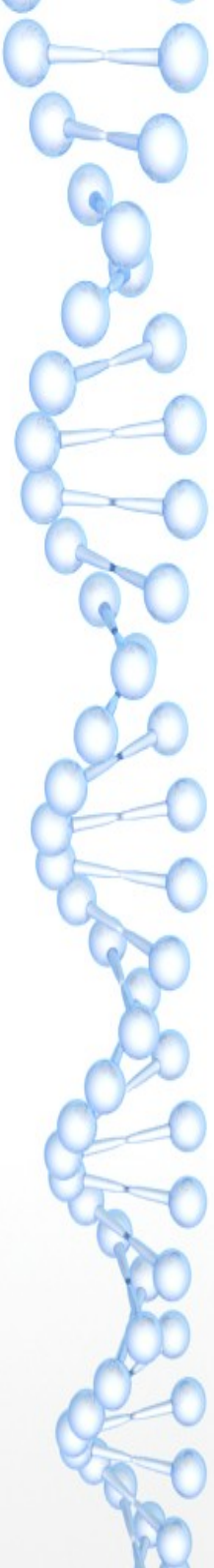
# Outline

- Introduction
- Related Work
- Knowledge Graphs
- Cross-Language Knowledge Graph Analysis
- Evaluation
- **Conclusions and future work**



# Conclusions

- Knowledge graphs...
  - enable language independence.





# Conclusions

- Knowledge graphs...
  - enable language independence.
  - can be used in cross-language plagiarism detection.



# Conclusions

- Knowledge graphs...
  - enable language independence.
  - can be used in cross-language plagiarism detection.
  - enable CL-KGA model to outperform the state-of-the-art.



## Future work (Ph.D.)

- We will investigate further how the task of cross-language plagiarism detection can be approached using multilingual semantic networks.



## Future work (Ph.D.)

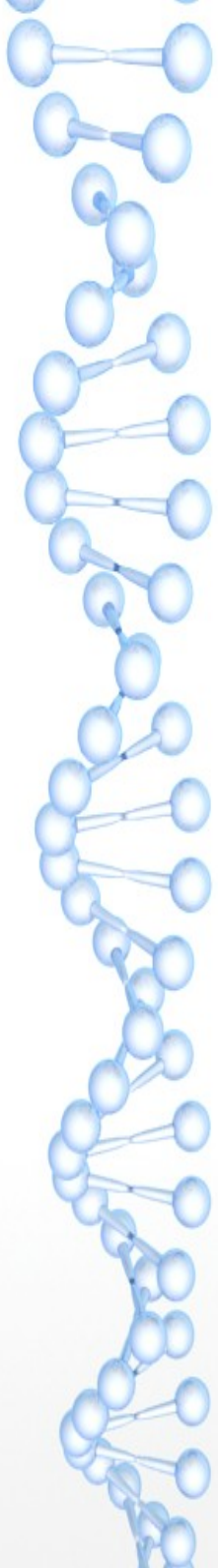
- We will investigate further how the task of cross-language plagiarism detection can be approached using multilingual semantic networks.
- We will study the possible use of knowledge graphs to perform other tasks such as:
  - Monolingual and cross-lingual similarity analysis
  - Cross-language document retrieval
  - Cross-language document categorization
  - Monolingual and cross-lingual domain adaptation

# Publications

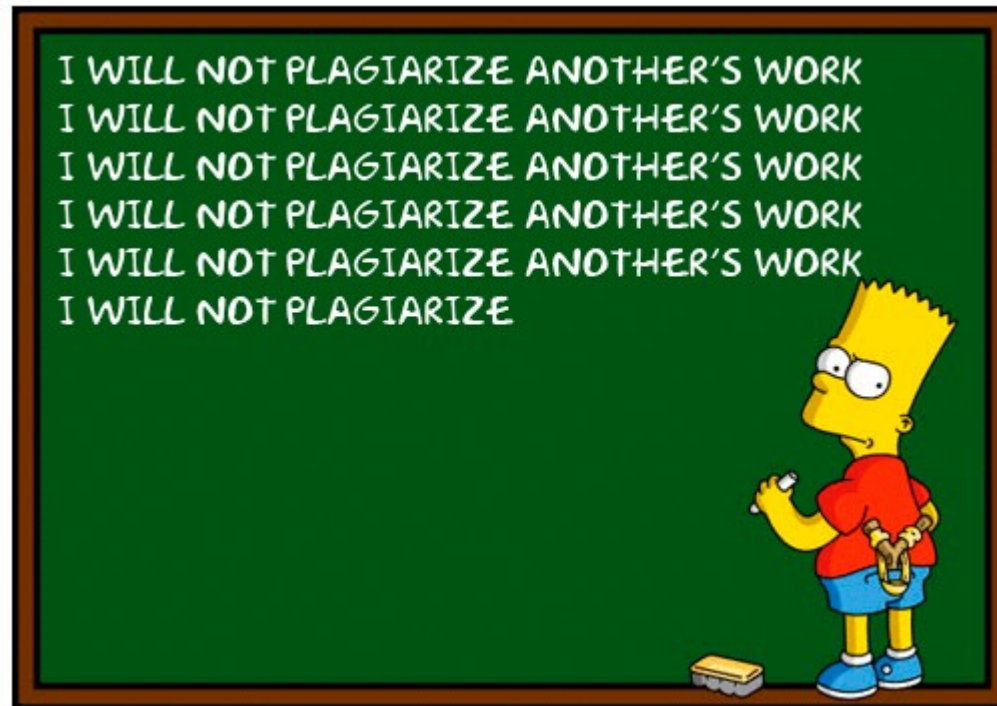
- Franco-Salvador M., Gupta P., Rosso P. Cross-Language Plagiarism Detection Using a Multilingual Semantic Network. In 35th European Conference on Information Retrieval (ECIR'13). Springer-Verlag, LNCS(7814), Moscow, Russia, 2013.
- Franco-Salvador M., Gupta P., Rosso P. Knowledge Graphs as Context Models: Improving the Detection of Cross-Language Plagiarism with Paraphrasing. In Proc. of the PROMISE Winter School 2013, Bressanone, Italy, 2013.
- Franco-Salvador M., Gupta P., Rosso P. Análisis de similitud basado en grafos: una nueva aproximación a la detección de plagio translingüe. Sociedad Española de Procesamiento del Lenguaje Natural (SEPLN) , ISSN 1135-5948, num. 50, 2013.
- Franco-Salvador M., Gupta P., Rosso P. Detección de plagio translingüe utilizando el diccionario estadístico de BabelNet. Computacion y Sistemas, Revista Iberoamericana de Computación, ISSN 1405-5546, vol. 16, num. 4, pp. 383-390, 2012.

Thanks to





Thanks for your time :)







# References

- Mcnamee, P. and Mayfield, J. Character n-gram tokenization for European language text retrieval. *In Information Retrieval*, 7(1):73–97, 2004.
- Barrón-Cedeño, A., Rosso, P., Pinto, D., and Juan, A. On cross-lingual plagiarism analysis using a statistical model. *In Proc. of the ECAI'08 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse, PAN'08*, 2008.
- Potthast, M., Barrón-Cedeño, A., Stein, B., Rosso, P. Cross-Language Plagiarism Detection. *In: Languages Resources and Evaluation. Special Issue on Plagiarism and Authorship Analysis*, vol. 45, num. 1, pp.45-62, 2011.
- Potthast, M., Eiselt, A., Barrón-Cedeño, A., Stein B. and Rosso P. Overview of the 3rd International Competition on Plagiarism Detection. *In: Petras V., Forner P., Clough P. (Eds.), Notebook Papers of CLEF 2011 LABs and Workshops, CLEF-2011, Amsterdam, The Netherlands, September 19-22, 2011.*
- Navigli, R. and Ponzetto, S. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence*, 193, Elsevier, pp. 217-250, 2012.
- Gupta P., Barrón-Cedeño A. and Rosso P. Cross-language High Similarity Search using a Conceptual Thesaurus. *In Proc. of CLEF 2012 (Rome, Italy)*, 2012
- Barrón-Cedeño, A., Gupta, P. and Rosso, P. Methods for Cross-Language Plagiarism Detection. *In: Knowledge-Based Systems. Volume 50*, pp. 211–217, 2013.





## Appendix: Detailed analysis

1: **Given**  $d$  and  $D'$ :

// **Detailed analysis**

2:  $S \leftarrow \{split(d, w, l)\}$      $S' \leftarrow \{split(d', w, l)\}$

3: for every  $s \in S$ :

4:     $P_{s,s'} \leftarrow \{argmax_{s' \in S'}^{5} sim(s, s')\}$

// **Post-processing**

5: until no change:

6:    for every combination of pairs  $p \in P_{s,s'}$ :

7:        if  $\delta(p_i, p_j) < thres_1$ :

8:             $merge\_fragments(p_i, p_j)$

// **Output**

9: return  $\{p \in P_{s,s'} / |p| > thres_2\}$