Using Word Embedding for Cross-Language Plagiarism Detection

Compilatio.net IT SOLUTION FOR PLAGIARISM DETECTION

UNIVERSITÉ Grenoble Alpes

LIG

Authors

Jérémy Ferrero **Frédéric Agnès** **Laurent Besacier** **Didier Schwab**

What is Cross-Language Plagiarism Detection?

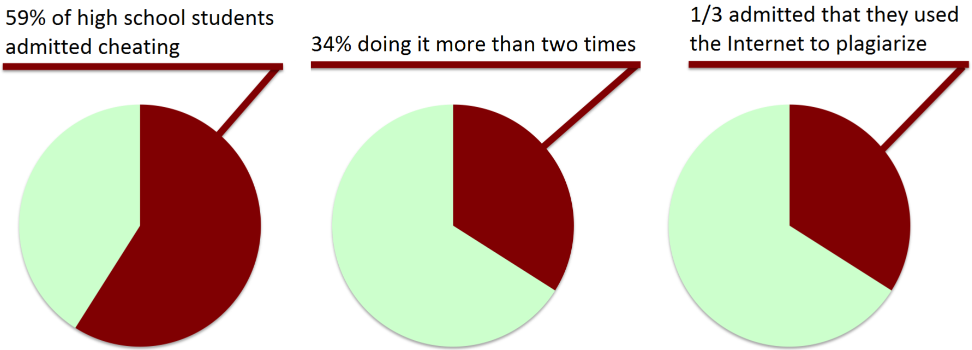
Cross-Language Plagiarism is a plagiarism by translation, *i.e.* a text has been plagiarized while being translated (manually or automatically).

présentation d'un tel log qui soit à la fois concise et exploitable. L'idée de base est qu'une requête résume une autre requête et qu'un log, qui est une séquence de requêtes, résume un autre log. Nous proposons également plusieurs stratégies

for summarizing and querying OLAP query logs. The basic idea is that a query summarizes another query and that a log, which is a sequence of queries, summarizes another log. Our formal framework includes a language to declaratively specify a

From a text in a language L, we must find similar passage(s) in other text(s) from among a set of candidate texts in language L' (cross-language textual similarity).

Why is it so important?



Sources:
- McCabe, D. (2010). Students' cheating takes a high-tech turn. In Rutgers Business School.
- Josephson Institute. (2011). What would honest Abe Lincoln say?

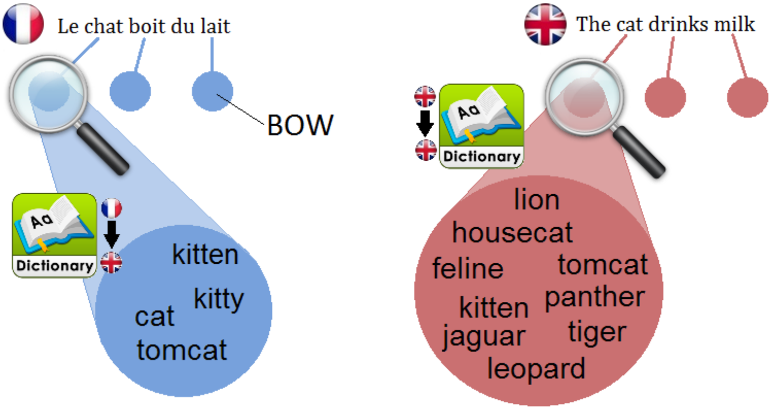
Research Questions

- Are Word Embeddings useful for cross-language plagiarism detection?
- Is syntax weighting in distributed representations of sentences useful for the text entailment?
- Are cross-language plagiarism detection methods complementary?

State-of-the-Art Methods

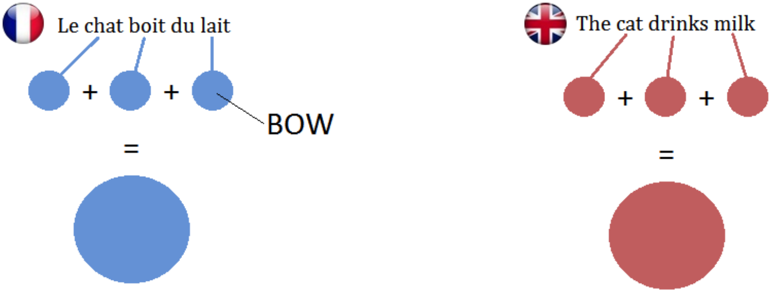
- Syntax-Based Models
  Length Model, **CL-C*n*G** [Mcnamee and Mayfield, 2004, Potthast et al., 2011], Cognateness
- Dictionary-Based Models
  CL-VSM, **CL-CTS** [Gupta et al., 2012, Pataki, 2012]
- Parallel Corpora-Based Models
  CL-ASA [Barrón-Cedeño et al., 2008, Pinto et al., 2009], CL-LSI, CL-KCCA
- Comparable Corpora-Based Models
  CL-KGA, **CL-ESA** [Gabrilovich and Markovitch, 2007, Potthast et al., 2008]
- MT-Based Models
  Translation + Monolingual Analysis [Muhr et al., 2010, Barrón-Cedeño, 2012]

Augmented CL-CTS

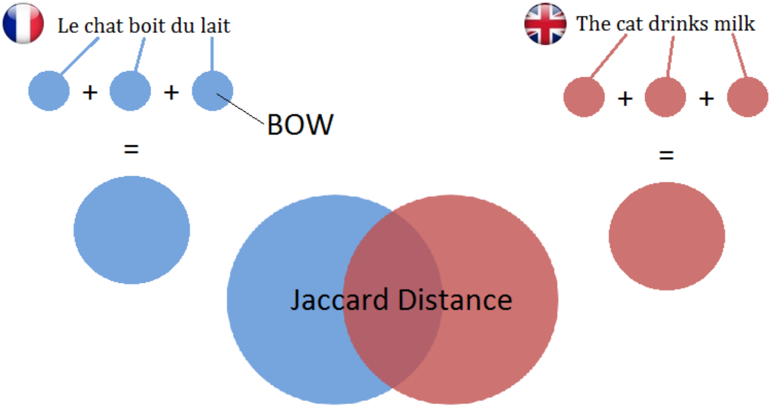


We use DBNary [Sérasset, 2015] as linked lexical resource.

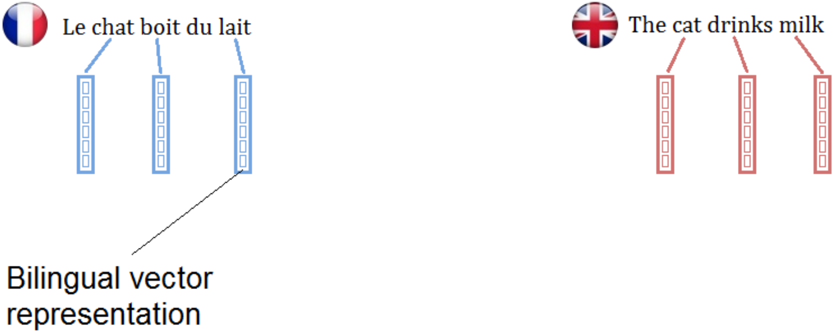
Augmented CL-CTS



Augmented CL-CTS

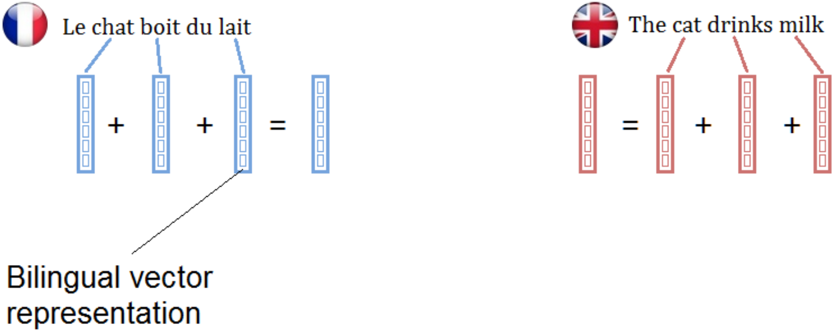


CL-WES: Cross-Language Word Embedding-based Similarity



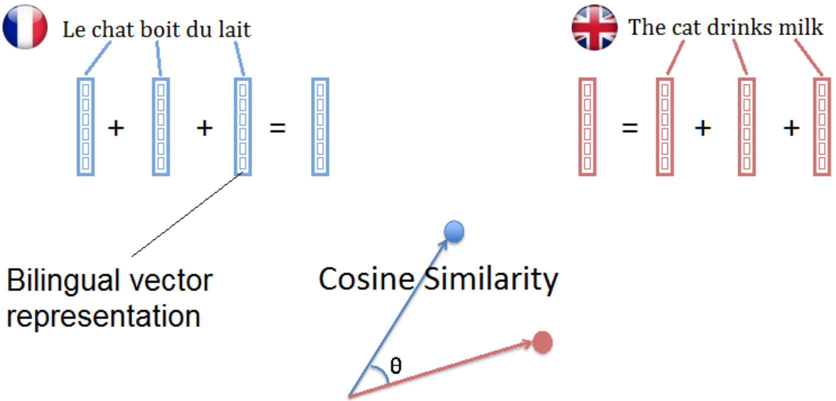
This feature is available in MultiVec [Berard et al., 2016] (`https://github.com/eske/multivec`)

CL-WES: Cross-Language Word Embedding-based Similarity



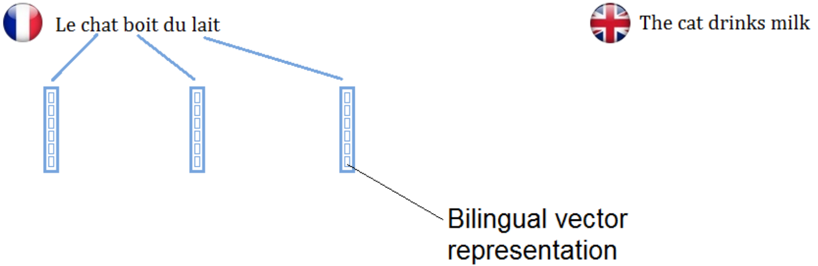
This feature is available in MultiVec [Berard et al., 2016] (https://github.com/eske/multivec)

CL-WES: Cross-Language Word Embedding-based Similarity



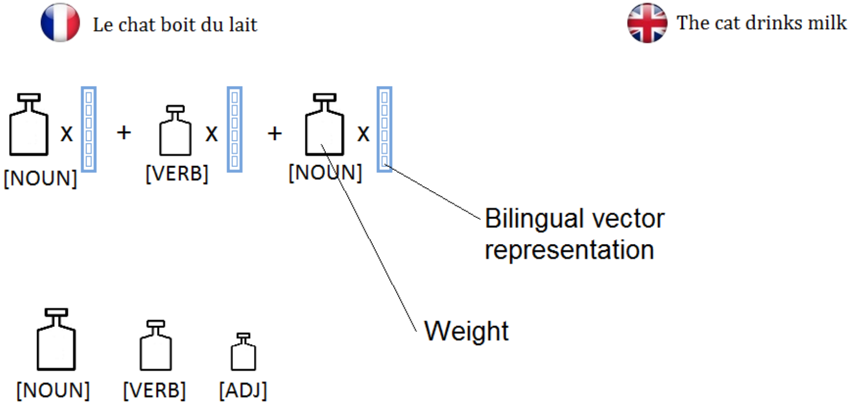
This feature is available in MultiVec [Berard et al., 2016] (`https://github.com/eske/multivec`)

CL-WESS: Cross-Language Word Embedding-based Syntax Similarity



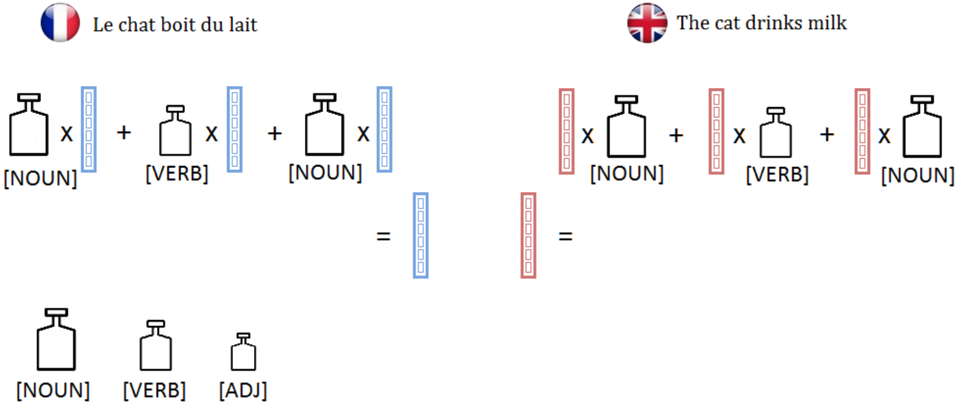
This feature is available in MultiVec [Berard et al., 2016] (`https://github.com/eske/multivec`)

CL-WESS: Cross-Language Word Embedding-based Syntax Similarity



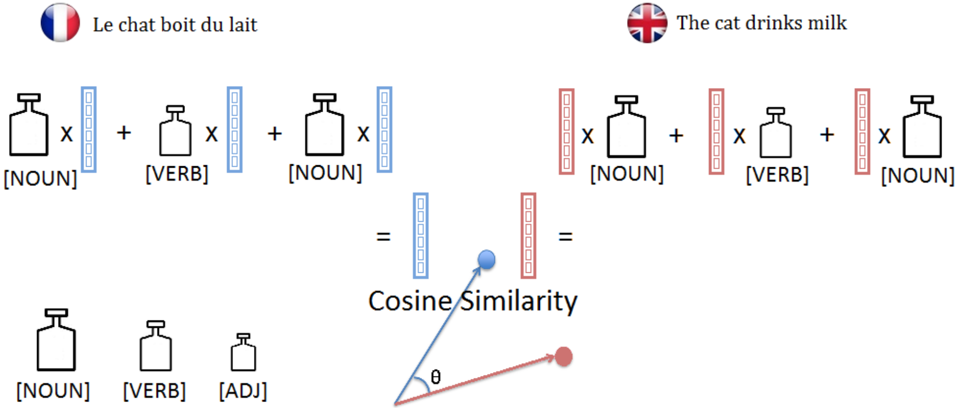
This feature is available in MultiVec [Berard et al., 2016] (`https://github.com/eske/multivec`)

CL-WESS: Cross-Language Word Embedding-based Syntax Similarity



This feature is available in MultiVec [Berard et al., 2016] (`https://github.com/eske/multivec`)

CL-WESS: Cross-Language Word Embedding-based Syntax Similarity



This feature is available in MultiVec [Berard et al., 2016] (`https://github.com/eske/multivec`)

Evaluation Dataset

[Ferrero et al., 2016][1]

- **French**, **English** and **Spanish**;
- **Parallel** and **comparable** (mix of Wikipedia, conference papers, product reviews, Europarl and JRC);
- Different granularities: **document** level, **sentence** level and **chunk** level;
- **Human** and **machine translated** texts;
- **Obfuscated** (to make the similarity detection more complicated) and **without added noise**;
- Written and translated by **multiple types of authors**;
- Cover **various fields**.

[1] A Multilingual, Multi-style and Multi-granularity Dataset for Cross-language Textual Similarity Detection. In Proceedings of LREC 2016.
`https://github.com/FerreroJeremy/Cross-Language-Dataset`

Evaluation Protocol

- We compared each English textual unit to its corresponding French unit and to 999 other units randomly selected;
- We threshold the obtained distance matrix to find the threshold giving the best F_1 score;
- We repeat these two steps 10 times, leading to a 10 folds:
  - 2 folds for tuning (CL-WESS) and fusion (Decision Tree)
  - 8 folds for validation

Results

Overall (%)		
	Chunk-Level	Sentence-Level
State-of-the-Art Methods		
CL-C3G	50.76	49.34
CL-CTS	**42.84**	**47.50**
CL-ASA	47.32	35.81
CL-ESA	14.81	14.44
T+MA	37.12	37.42
New Proposed Methods		
CL-CTS-WE	**46.67**	**50.69**
CL-WES	41.95	41.43
CL-WESS	53.73	56.35
Decision Tree	89.15	88.50

Table: Average F_1 scores of methods applied on EN$\rightarrow$FR sub-corpora.

- **CL-CTS-WE boosts CL-CTS (+3.83% on chunks and +3.19% on sentences);**
- CL-WESS boosts CL-WES (+11.78% on chunks and +14.92% on sentences);
- CL-WESS is better than CL-C3G (+2.97% on chunks and +7.01% on sentences);
- Decision Tree fusion significantly improves the results.

CL-CTS-WE: Cross-Language Conceptual Thesaurus-based Similarity with Word-Embedding

Results

	Overall (%)	
	Chunk-Level	Sentence-Level
State-of-the-Art Methods		
CL-C3G	50.76	49.34
CL-CTS	42.84	47.50
CL-ASA	47.32	35.81
CL-ESA	14.81	14.44
T+MA	37.12	37.42
New Proposed Methods		
CL-CTS-WE	46.67	50.69
CL-WES	**41.95**	**41.43**
CL-WESS	**53.73**	**56.35**
Decision Tree	89.15	88.50

Table: Average F_1 scores of methods applied on EN→FR sub-corpora.

- CL-CTS-WE boosts CL-CTS (+3.83% on chunks and +3.19% on sentences);
- **CL-WESS boosts CL-WES (+11.78% on chunks and +14.92% on sentences);**
- CL-WESS is better than CL-C3G (+2.97% on chunks and +7.01% on sentences);
- Decision Tree fusion significantly improves the results.

CL-WES: Cross-Language Word Embedding-based Similarity

CL-WESS: Cross-Language Word Embedding-based Syntax Similarity

Results

	Overall (%)	
	Chunk-Level	**Sentence-Level**
	State-of-the-Art Methods	
CL-C3G	**50.76**	**49.34**
CL-CTS	42.84	47.50
CL-ASA	47.32	35.81
CL-ESA	14.81	14.44
T+MA	37.12	37.42
	New Proposed Methods	
CL-CTS-WE	46.67	50.69
CL-WES	41.95	41.43
CL-WESS	**53.73**	**56.35**
Decision Tree	89.15	88.50

Table: Average F_1 scores of methods applied on EN→FR sub-corpora.

- CL-CTS-WE boosts CL-CTS (+3.83% on chunks and +3.19% on sentences);
- CL-WESS boosts CL-WES (+11.78% on chunks and +14.92% on sentences);
- **CL-WESS is better than CL-C3G (+2.97% on chunks and +7.01% on sentences);**
- Decision Tree fusion significantly improves the results.

CL-C3G: Cross-Language Character 3-Gram

CL-WESS: Cross-Language Word Embedding-based Syntax Similarity

Results

	Overall (%)	
	Chunk-Level	**Sentence-Level**
	State-of-the-Art Methods	
CL-C3G	50.76	49.34
CL-CTS	42.84	47.50
CL-ASA	47.32	35.81
CL-ESA	14.81	14.44
T+MA	37.12	37.42
	New Proposed Methods	
CL-CTS-WE	46.67	50.69
CL-WES	41.95	41.43
CL-WESS	53.73	56.35
Decision Tree	**89.15**	**88.50**

Table: Average F_1 scores of methods applied on EN→FR sub-corpora.

- CL-CTS-WE boosts CL-CTS (+3.83% on chunks and +3.19% on sentences);
- CL-WESS boosts CL-WES (+11.78% on chunks and +14.92% on sentences);
- CL-WESS is better than CL-C3G (+2.97% on chunks and +7.01% on sentences);
- **Decision Tree fusion significantly improves the results.**

Decision Tree fusion: C4.5 [Quinlan, 1993]

Conclusion

- **Augmentation of several baseline approaches** using word embeddings instead of lexical resources;
- **CL-WESS** beats in overall the precedent best state-of-the-art methods;
- **Methods are complementary** and their fusion significantly helps cross-language textual similarity detection performance;
- **Winning method at SemEval-2017 Task 1 track 4a**, *i.e.* the task on Spanish-English Cross-lingual Semantic Textual Similarity detection.

Thank you for your attention.
Do you have any questions?

jeremy.ferrero@compilatio.net
@FerreroJeremy
github.com/FerreroJeremy
fr.linkedin.com/in/FerreroJeremy
researchgate.net/profile/Jeremy_Ferrero

Augmented CL-CTS

- *CL-CTS-WE* uses the top 10 closest words in the embeddings model to build the BOW of a word;
- A BOW of a sentence is a merge of the BOW of its words;
- Jaccard distance between the two BOW.

CL-WES: Cross-Language Word Embedding-based Similarity

The similarity between two sentences S and S' is calculated by Cosine Distance between the two vectors V and V', built such as:

$$V(S) = \sum_{i=1}^{|S|} (vector(u_i)) \tag{1}$$

- u_i is the i^{th} word of S;
- *vector* is the function which gives the word embedding vector of a word.

This feature is available in *MultiVec*[2] [Berard et al., 2016].

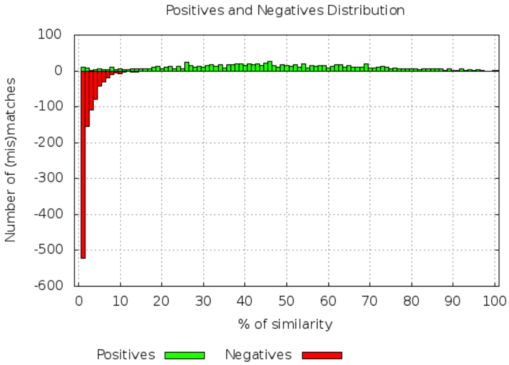
[2] `https://github.com/eske/multivec`

CL-WESS: Cross-Language Word Embedding-based Syntax Similarity

$$V(S) = \sum_{i=1}^{|S|} (weight(pos(u_i)).vector(u_i)) \quad (2)$$

- u_i is the i^{th} word of S;
- *pos* is the function which gives the universal part-of-speech tag of a word;
- *weight* is the function which gives the weight of a part-of-speech;
- *vector* is the function which gives the word embedding vector of a word;
- . is the scalar product.

Complementarity



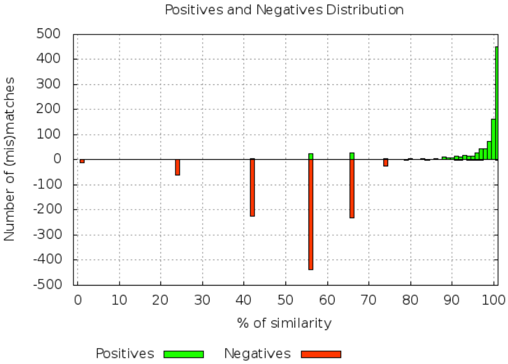
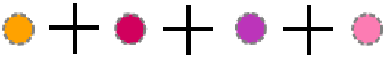


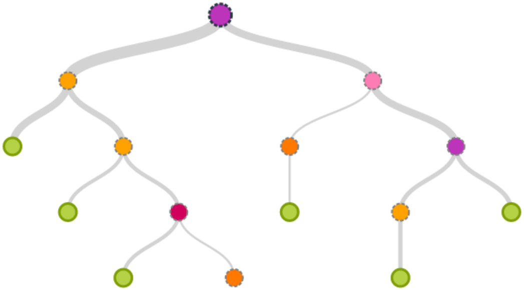
Figure: Distribution histograms of *CL-CNG* (left) and *CL-ASA* (right) for 1000 positives and 1000 negatives (mis)matches.

Fusions

Weighted Average Fusion



Decision Tree Fusion C4.5 [Quinlan, 1993]



Weighted Fusion

$$fus(M) = \frac{\sum_{j=1}^{|M|}(w_j * m_j)}{\sum_{j=1}^{|M|} w_j} \quad (3)$$

- M is the set of the scores of the methods for one match;
- m_j and w_j are the score and the weight of the j^{th} method respectively.

Results at Chunk-Level

Chunk level						
Methods	**Wikipedia (%)**	**TALN (%)**	**JRC (%)**	**APR (%)**	**Europarl (%)**	**Overall (%)**
CL-C3G	63.04	**40.80**	**36.80**	**80.69**	53.26	**50.76**
CL-CTS	58.05	33.66	30.15	67.88	45.31	42.84
CL-ASA	23.70	23.24	33.06	26.34	**55.45**	47.32
CL-ESA	**64.86**	23.73	13.91	23.01	13.98	14.81
T+MA	58.26	38.90	28.81	73.25	36.60	37.12
CL-CTS-WE	**58.00**	**38.04**	31.73	**73.13**	49.91	46.67
CL-WES	37.53	21.70	32.96	39.14	46.01	41.95
CL-WESS	52.68	34.49	**45.00**	56.83	**57.06**	**53.73**
Average fusion	81.34	65.78	61.87	91.87	79.77	75.82
Weighed fusion	84.61	69.69	67.02	94.38	83.74	80.01
Decision Tree	**95.25**	**74.10**	**72.19**	**97.05**	**95.16**	**89.15**

Table: Average F_1 scores of cross-language similarity detection methods applied on chunk-level EN→FR sub-corpora – 8 folds validation.

Results at Sentence-Level

Sentence level						
Methods	**Wikipedia (%)**	**TALN (%)**	**JRC (%)**	**APR (%)**	**Europarl (%)**	**Overall (%)**
CL-C3G	48.24	**48.19**	**36.85**	61.30	52.70	**49.34**
CL-CTS	46.71	38.93	28.38	51.43	**53.35**	47.50
CL-ASA	27.68	27.33	34.78	25.95	36.73	35.81
CL-ESA	**50.89**	14.41	14.45	14.18	14.09	14.44
T+MA	50.39	37.66	32.31	**61.95**	37.70	37.42
CL-CTS-WE	**47.26**	**43.93**	31.63	57.85	56.39	50.69
CL-WES	28.48	24.37	33.99	39.10	44.06	41.43
CL-WESS	45.65	40.45	**48.64**	**58.08**	**58.84**	**56.35**
Decision Tree	**80.45**	**80.89**	**72.70**	**78.91**	**94.04**	**88.50**

Table: Average F_1 scores of cross-language similarity detection methods applied on sentence-level EN→FR sub-corpora – 8 folds validation.

References I

Barrón-Cedeño, A. (2012).
On the Mono- and Cross-Language Detection of Text Re-Use and Plagiarism.
In *PhD thesis*, València, Spain.

Barrón-Cedeño, A., Rosso, P., Pinto, D., and Juan, A. (2008).
On Cross-lingual Plagiarism Analysis using a Statistical Model.
In Benno Stein and Efstathios Stamatatos and Moshe Koppel, editor, *Proceedings of the ECAI'08 PAN Workshop: Uncovering Plagiarism, Authorship and Social Software Misuse*, pages 9–13, Patras, Greece.

References II

Berard, A., Servan, C., Pietquin, O., and Besacier, L. (2016).
MultiVec: a Multilingual and Multilevel Representation Learning Toolkit for NLP.
In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4188–4192, Portoroz, Slovenia. European Language Resources Association (ELRA).

Ferrero, J., Agnès, F., Besacier, L., and Schwab, D. (2016).
A Multilingual, Multi-style and Multi-granularity Dataset for Cross-language Textual Similarity Detection.
In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4162–4169, Portoroz, Slovenia. European Language Resources Association (ELRA).

References III

Gabrilovich, E. and Markovitch, S. (2007).
Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis.
In *Proceedings of the 20th International Joint Conference on Artifical Intelligence (IJCAI'07)*, pages 1606–1611, Hyderabad, India. Morgan Kaufmann Publishers Inc.

Gupta, P., Barrón-Cedeño, A., and Rosso, P. (2012).
Cross-language High Similarity Search using a Conceptual Thesaurus.
In *Information Access Evaluation. Multilinguality, Multimodality, and Visual Analytics*, pages 67–75, Rome, Italy. Springer Berlin Heidelberg.

References IV

Mcnamee, P. and Mayfield, J. (2004).
Character N-Gram Tokenization for European Language Text Retrieval.
In *Information Retrieval Proceedings*, volume 7, pages 73–97. Kluwer Academic Publishers.

Muhr, M., Kern, R., Zechner, M., and Granitzer, M. (2010).
External and Intrinsic Plagiarism Detection Using a Cross-Lingual Retrieval and Segmentation System - Lab Report for PAN at CLEF 2010.
In Braschler, M., Harman, D., and Pianta, E., editors, *CLEF Notebook*, Padua, Italy.

References V

Pataki, M. (2012).
A New Approach for Searching Translated Plagiarism.
In *Proceedings of the 5th International Plagiarism Conference*, pages 49–64, Newcastle, UK.

Pinto, D., Civera, J., Juan, A., Rosso, P., and Barrón-Cedeño, A. (2009).
A Statistical Approach to Crosslingual Natural Language Tasks.
In *CEUR Workshop Proceedings*, volume 64 of *Journal of Algorithms*, pages 51–60.

Potthast, M., Barrón-Cedeño, A., Stein, B., and Rosso, P. (2011).
Cross-Language Plagiarism Detection.
In *Language Resources and Evaluation*, volume 45, pages 45–62.

References VI

Potthast, M., Stein, B., and Anderka, M. (2008).
A Wikipedia-Based Multilingual Retrieval Model.
In *30th European Conference on IR Research (ECIR'08)*, volume 4956 of *LNCS of Lecture Notes in Computer Science*, pages 522–530, Glasgow, Scotland. Springer.

Quinlan, J. R. (1993).
C4.5: Programs for Machine Learning.
The Morgan Kaufmann series in machine learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

References VII

Sérasset, G. (2015).
DBnary: Wiktionary as a Lemon-Based Multilingual Lexical Resource in RDF.
In *Semantic Web Journal (special issue on Multilingual Linked Open Data)*, volume 6, pages 355–361.