A Multilingual, Multi-Style and Multi-Granularity Dataset for Cross-Language Textual Similarity Detection

Jérémy Ferrero[1,2], Frédéric Agnès[1], Laurent Besacier[2], Didier Schwab[2]

[1] Compilatio, 276 rue du Mont Blanc, 74540 Saint-Félix, France
[2] LIG-GETALP, Univ. Grenoble Alpes, France

Cross-Language Plagiarism

Plagiarism is an act of fraud...

- ...to steal and pass off an idea as one's own...
- ...without the consent of the author...
- ...and without crediting the source. [1]

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

The challenge in detecting this kind of plagiarism is that the suspicious document is in a language different from its source.

> 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

References

[1] Website. http://www.plagiarism.org/plagiarism-101/what-is-plagiarism.

[2] Martin Potthast *et al.* Cross-Language Plagiarism Detection. In *Language Ressources and Evaluation*, volume 45, pages 45–62, 2011.

[3] Ralf Steinberger. JRC-ACQUIS Multilingual Parallel Corpus, 2011. European Commission's Joint Research Centre. version 3.0 ISLRN: 821-325-977-001-1.

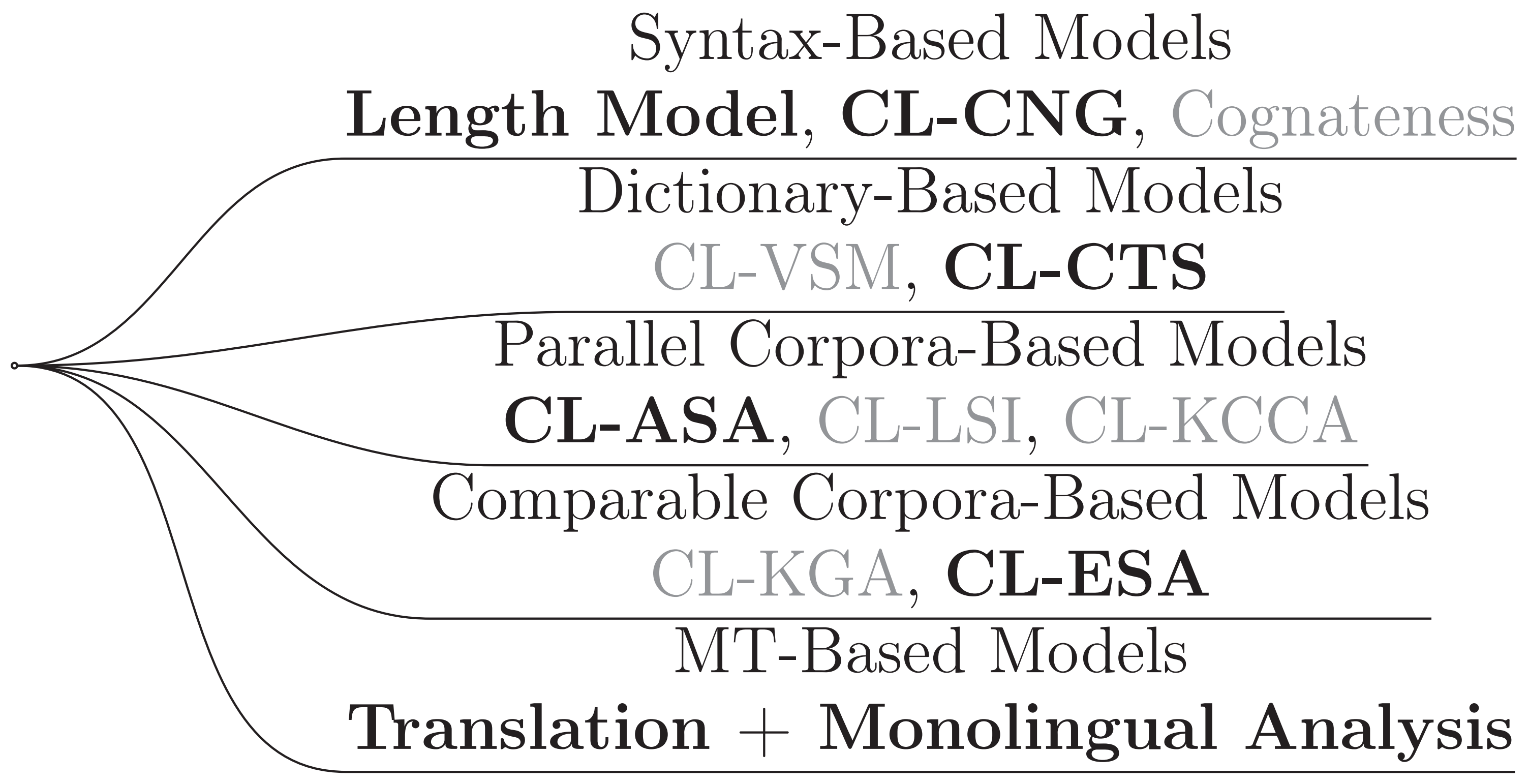
[4] Philipp Koehn. Europarl: European Parliament Proceedings Parallel Corpus, 2005. European Language Resources Association. version 6.0.

[5] Martin Potthast *et al.* Wikipedia Corpus, 2011.

[6] Martin Potthast *et al.* PAN-PC-11 corpus, 2010. Bauhaus-Universität Weimar & Universidad Politécnica de Valencia.

[7] Peter Prettenhofer and Benno Stein. Webis-CLS-10: Cross-Lingual Sentiment Dataset, 2010. version 1.0 (11.5.2010).

Textual Similarity Detection Methods



Syntax-Based Models
Length Model, **CL-CNG**, Cognateness
Dictionary-Based Models
CL-VSM, **CL-CTS**
Parallel Corpora-Based Models
CL-ASA, CL-LSI, CL-KCCA
Comparable Corpora-Based Models
CL-KGA, **CL-ESA**
MT-Based Models
Translation + Monolingual Analysis

Taxonomy of different approaches of cross-language similarity detection methodologies [2] (in bold, the methods that we have evaluated on our dataset).

Our Dataset for Cross-Language Textual Similarity Detection

Our dataset is composed of texts:

- in **French**, **English** and **Spanish**;
- aligned at the **document-**, **sentence-** and **chunk- level**;
- aligned from **parallel or comparable** collections;
- covering various fields;
- **translated** by humans (professionals or not) or automatically;
- **altered** or without added noise.

Sub-corpus	Languages	# Documents	# Sentences	# Chunks
JRC-Acquis [3]	EN, FR, ES	≃10,000	≃150,000	≃10,000
Europarl [4]	EN, FR, ES	≃10,000	≃475,000	≃25,600
Wikipedia [5]	EN, FR, ES	≃10,000	≃5,000	≃150
PAN-PC-11 [6]	EN, ES	≃3,000	≃90,000	≃1,400
APR [7]	EN, FR	≃6,000	≃25,000	≃2,600
Conference papers	EN, FR	≃35	≃1,300	≃300

Table 1: Statistics of our dataset.

The entire dataset is made available to the community on GitHub: **https://github.com/FerreroJeremy/Cross-Language-Dataset**.

Results

Methods	Wiki (%)	Conf. papers (%)	JRC (%)	APR (%)	Europarl (%)	Overall (%)
Random Baseline	00.21 ±0.019	00.22 ±0.025	00.23 ±0.029	00.22 ±0.025	00.24 ±0.030	00.22
Length Model	00.30 ±0.000	00.30 ±0.000	00.30 ±0.000	00.30 ±0.000	00.30 ±0.000	00.30
CL-C3G	48.25 ±0.349	48.08 ±0.538	36.68 ±0.693	61.10 ±0.581	52.72 ±0.866	49.37
CL-CTS	46.68 ±0.437	38.67 ±0.552	28.21 ±0.612	50.82 ±0.687	53.21 ±0.601	43.52
CL-ASA	27.63 ±0.330	27.25 ±0.341	35.17 ±0.644	25.53 ±0.795	36.55 ±1.139	30.43
CL-ESA	51.14 ±0.875	14.25 ±0.334	14.44 ±0.341	13.93 ±0.714	13.91 ±0.618	21.53
T+MA	50.57 ±0.888	37.79 ±0.364	32.36 ±0.369	61.94 ±0.756	37.92 ±0.552	44.12
Average	44.85	33.21	29.37	42.66	38.86	

Table 2: Average F_1 scores and confidence intervals of state-of-the-art methods applied on the sentence-level EN-FR sub-corpora. The last row is the average F_1 scores from CL-C3G to T+MA.

- Random Baseline and Length Model show low performance;
- **CL-ESA seems to show better results on comparable corpora**, like Wikipedia;
- in contrast, **CL-ASA obtains better results on parallel corpora** such as JRC, Europarl or APR collections;
- **CL-C3G is in general the most effective method**, as long as the corpus includes named entities;
- right behind, **CL-CTS and T+MA are pretty efficient and versatile too**;
- CL-ESA is not very effective; it is the more time-consuming method and it is highly dependent on the corpus used.

There is a **strong correlation between the results of methods on the three granularities**.
The trend of the results on parallel corpora commonly used in evaluation tasks (e.g. JRC, Europarl) **correlate with the results on corpora from the target application domain** (e.g. scientific papers).

Perspectives

- Finalize the state-of-the-art evaluation with the other language pairs;
- **Boosting**: try a fusion of state-of-the-art cross-language textual similarity detection methods;
- Develop a **word embedding method for cross-language textual similarity detection**.

{jeremy, frederic}@compilatio.net, {jeremy.ferrero, laurent.besacier, didier.schwab}@imag.fr