CompiLIG at SemEval-2017 Task 1:
Cross-Language Plagiarism Detection Methods for Semantic Textual Similarity

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

IT SOLUTION FOR PLAGIARISM DETECTION

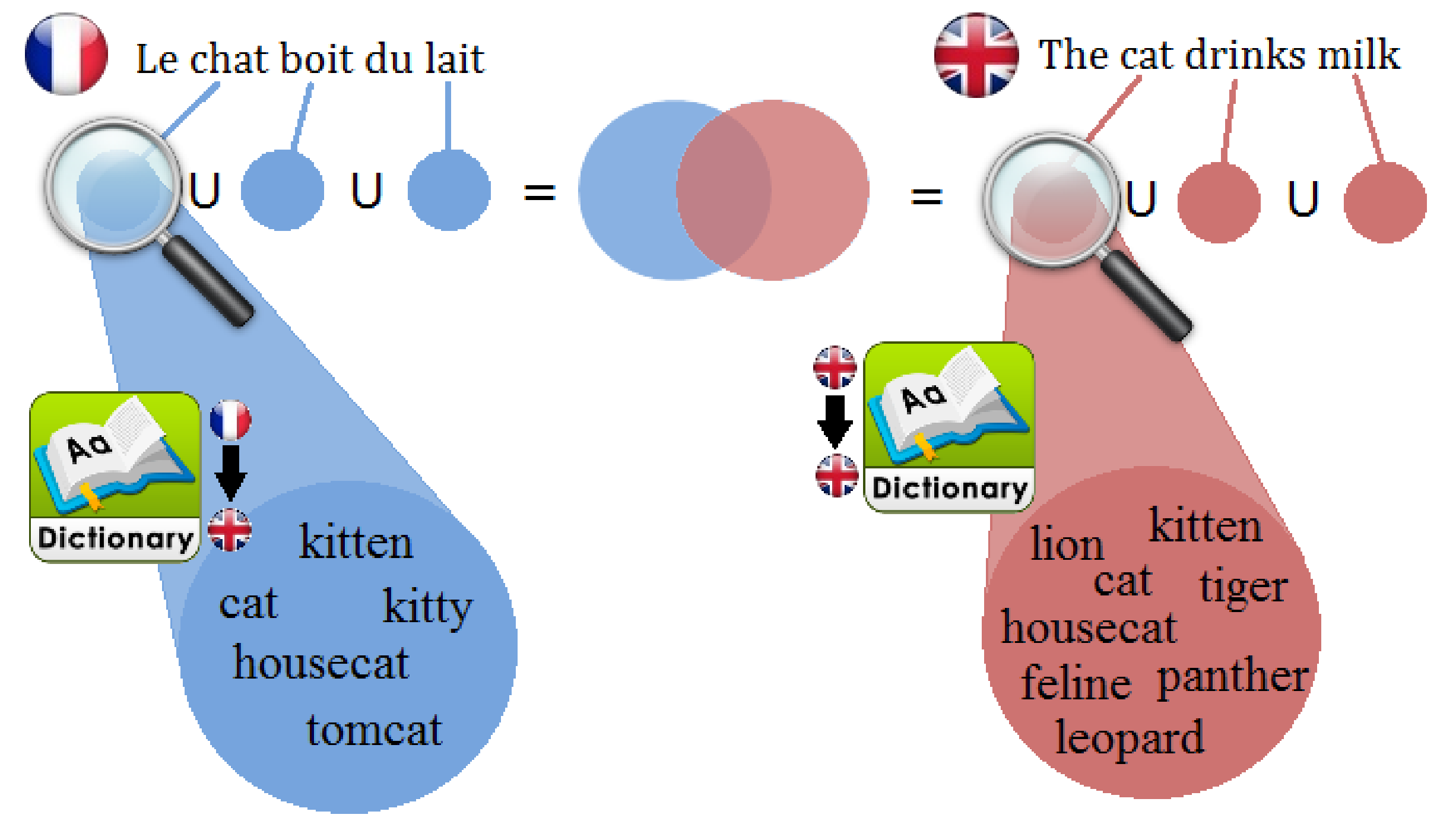
Abstract

We present our submitted systems for Semantic Textual Similarity (STS) Track 4 (**Spanish-English**) at SemEval-2017.

In our submission, we use syntax-based, dictionary-based, context-based, and MT-based methods. We also combine these methods in unsupervised and supervised way.

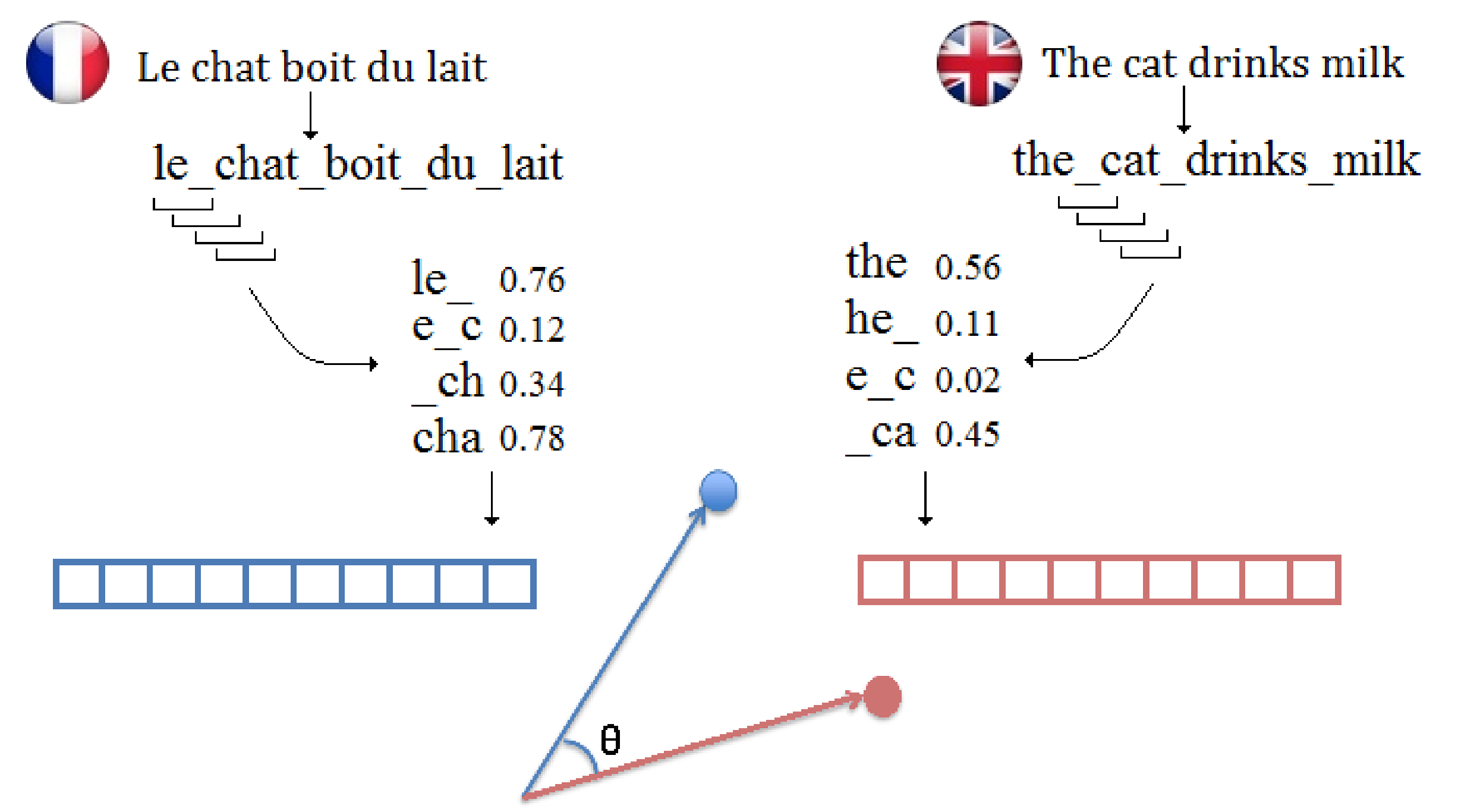
Our best run ranked **1st on track 4a** on 51 submitted systems, with a correlation of 83.02% with human annotations.

Conceptual Thesaurus Similarity (CTS)



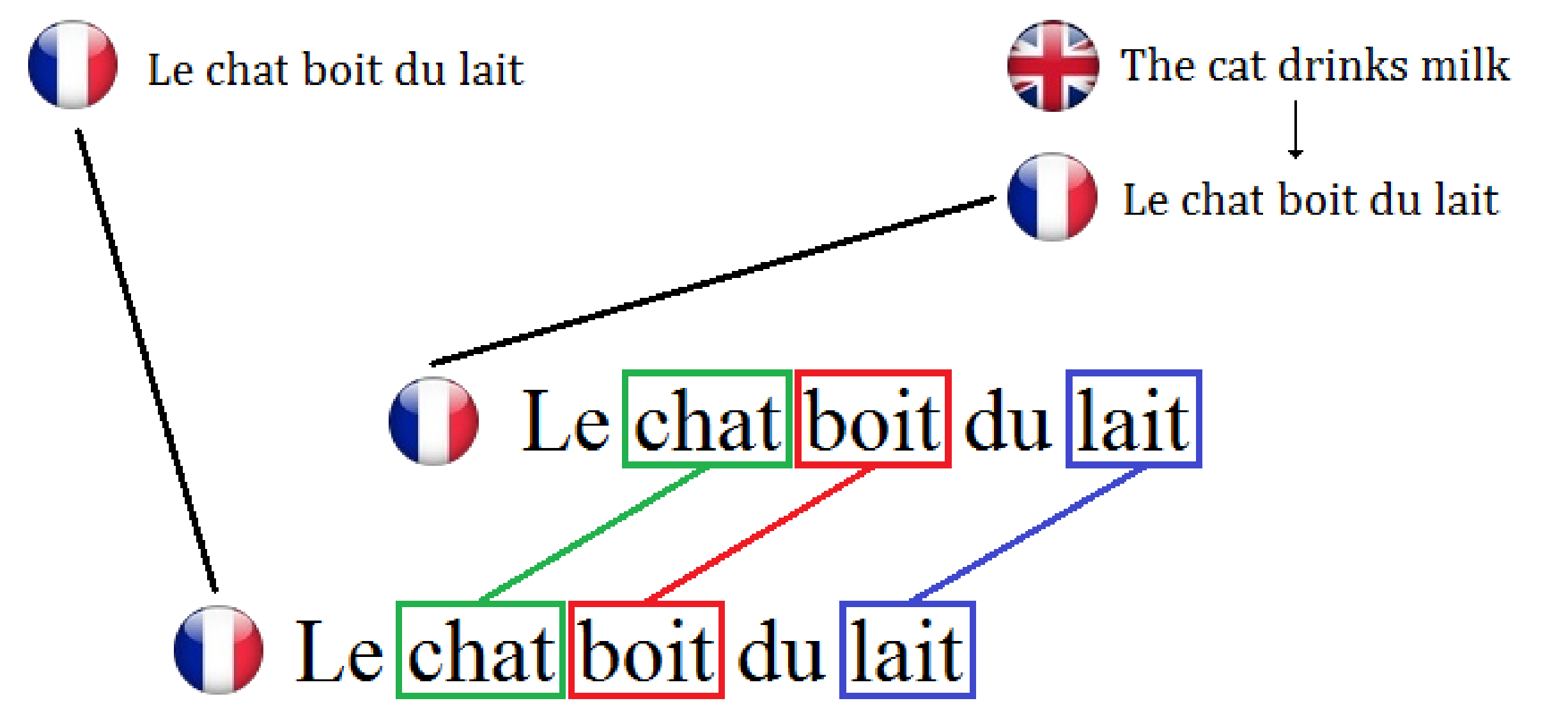
- Bag-of-words of a word = all its possible translations or nyms, jointly given by ontology DBNary [1] and by word embeddings with the MultiVec toolkit [2] ;
- **Bag-of-words of a sentence** = merge of the bag-of-words of its words ;
- **Syntactically** [3] and **frequentially** (*idf*) **weighted** augmentation of the **Jaccard distance** between the two built sentences bags.

Character 3-Gram (C3G)



- Only spaces and alphanumeric characters ;
- Segmention into 3-grams (sequences of 3 contiguous characters) ;
- Building of ***tf.idf* vectors of character 3-grams** ;
- **Cosine similarity** between the two vectors.

Translation + Word Alignment (T+WA)

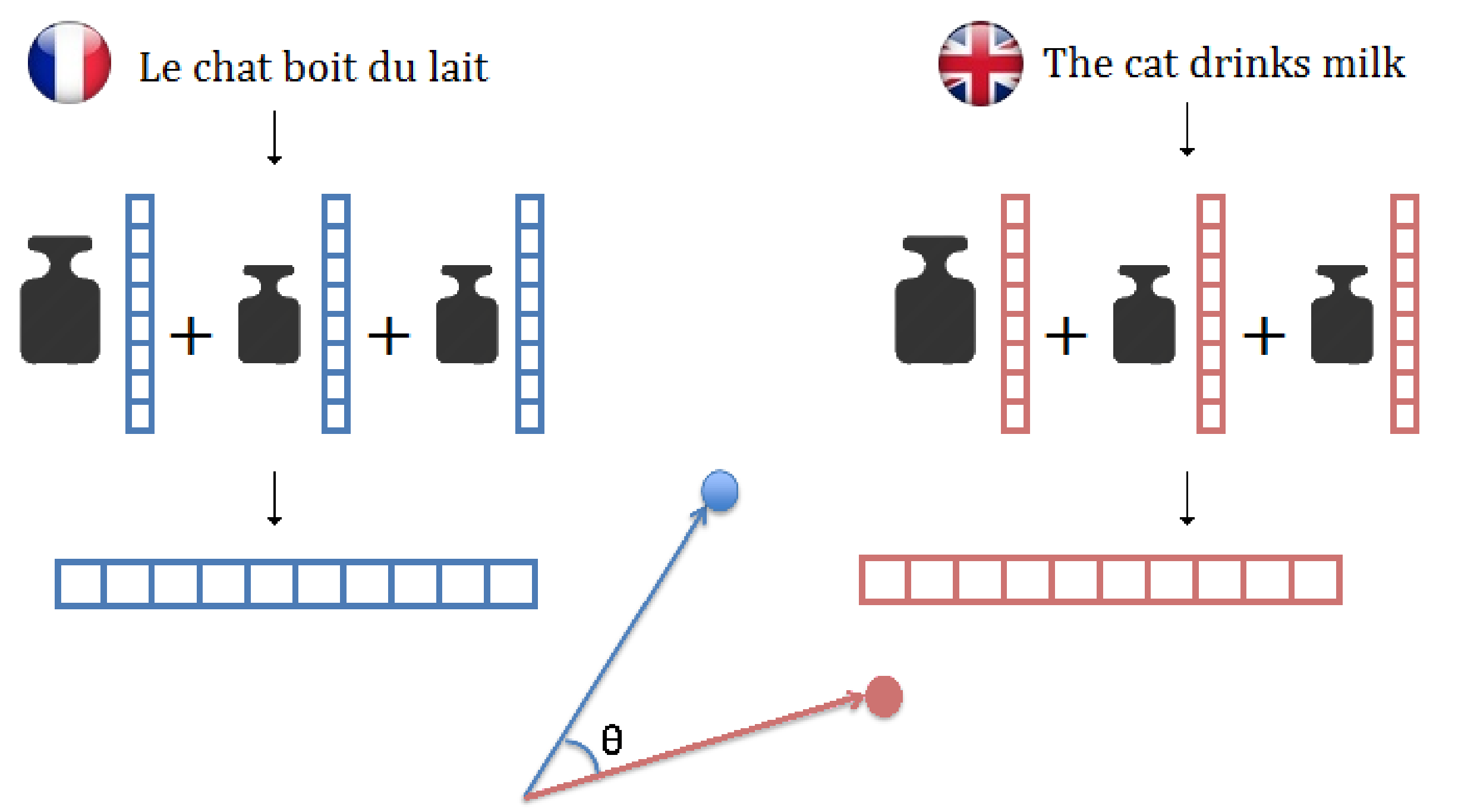


- **Translation** of the two sentences into the same language (Google Translate) ;
- **Monolingual Word Alignment** of UWB team [4] (winner of the SemEval-2016 cross-lingual STS task) ;
- **Frequentially** (*idf*) **weighted** augmentation of the **Jaccard distance**.

Publicly available on GitHub[a]!

[a] https://github.com/FerreroJeremy/monolingual-word-aligner

Word Embedding Similarity (WES)



- **Distributed representation of a sentence** = **syntactically** [3] and **frequentially** (*idf*) **weighted sum** of each **word vector** of this sentence ;
- **Cosine similarity** between the two vectors.

Publicly available through MultiVec toolkit[a] [2]!

[a] https://github.com/eske/multivec

Submissions & Results

Our three submissions are:

- Our **best method alone**: Cross-Language Conceptual Thesaurus-based Similarity (*CTS*) ;
- A **fusion by average** on *C3G*, *CTS* and *T+WA* ;
- A **M5′ model tree** [5] supervised fusion of our four presented methods.

Methods	SNLI (4a)	WMT (4b)	Mean
CTS	0.7684	0.1464	0.4574
Average	0.7910	0.1494	0.4702
M5′	**0.8302**	0.1550	0.4926

Table 1: Official results of SemEval-2017 STS track 4 evaluation.

- **1st** on 51 submitted systems, with 83.02% of correlation with human annotations, **on SNLI (track 4a)** ;
- Results on the WMT corpus (track 4b) are strangely low for all participating teams (see Discussion part).

Discussion

Methods	SNLI (4a)	WMT (4b)	Mean
Our Annotations			
CL-CTS	0.7981	0.5248	0.6614
Average	0.8105	0.4031	0.6068
M5′	0.8622	0.5374	0.6998
SemEval Gold Standard			
CL-CTS	0.8123	0.1739	0.4931
Average	0.8277	0.2209	0.5243
M5′	0.8536	0.1706	0.5121

Table 2: Results of our submitted systems scored on our 120 annotated pairs and on the same 120 SemEval (gold standard) annotated pairs.

- **Second annotator reference** ;
- Our methods behave the same way for both annotations on the SNLI corpus (4a) ;
- Huge difference on WMT corpus (4b) between our annotations and SemEval gold standard.

These results question the validity of the WMT corpus (4b).

References

[1] Gilles Sérasset. 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, 2015.

[2] Alexandre Berard, Christophe Servan, Olivier Pietquin, and Laurent Besacier. 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, May 2016. European Language Resources Association (ELRA).

[3] Jérémy Ferrero, Laurent Besacier, Didier Schwab, and Frédéric Agnès. Using Word Embedding for Cross-Language Plagiarism Detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, (EACL 2017)*, volume 2, pages 415–421, Valencia, Spain, April 2017. Association for Computational Linguistics.

[4] Tomas Brychcin and Lukas Svoboda. UWB at SemEval-2016 Task 1: Semantic textual similarity using lexical, syntactic, and semantic information. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, pages 588–594, San Diego, CA, USA, June 2016.

[5] Yong Wang and Ian H. Witten. Induction of model trees for predicting continuous classes. In *Proceedings of the poster papers of the European Conference on Machine Learning*, pages 128–137, Prague, Czech Republic, October 1997.