Team TMA at SemEval-2022 Task 8: Lightweight and Language-Agnostic News Similarity Classifier

Nicolas Stefanovitch
European Commission - Joint Research Centre
nicolas.stefanovitch@ec.europa.eu

Abstract

We present our contribution to the SemEval 22 Shared Task 8: *Multilingual news article similarity*. The approach is lightweight and language-agnostic, it is based on the computation of several lexicographic and embedding-based features, and the use of a simple ML approach: random forests. In a notable departure from the task formulation, which is a ranking task, we tackled this task as a classification one. We present a detailed analysis of the behaviour of our system under different settings.

1 Introduction

Detecting similar news is a key component of media monitoring, such as the one done by the Europe Media Monitor[1], which collects daily about half a million articles in more than 80 languages. This shared task (Chen et al., 2022) allows to study two important challenges that arise in practice: articles can be similar to several extent and in different ways; massive multilingualism requires language agnostic approaches. We report in this paper the result of an experimental approach to tackle this task.

We considered computing simple lexicographic and embedding-based features and using simple ML approach for complexity reasons, having in mind that pair-wise comparison of half a million articles each day is not possible with heavier solution without massive resource cost. There are two important tasks when building such large scale news clustering: determining which pair of articles are worth comparing, and computing the actual similarity between pairs of articles. While this shared task deals only with the second approach, the presented system has been designed to tackle also the first one.

[1] https://emm.newsbrief.eu/

2 System Description

Our system computes several lexicographic- and embedding-based similarity features, which are fed to a standard random forest. We used the following hyperparameters: 100 trees and a leaf split parameter of 3 data points. On the train set, these parameters avoided extreme overfitting to the data, while not degrading significantly the performances. We used 5-folds cross validation in order to choose the hyper parameters.

For each article we consider separately the similarities related to 4 fields of the news item: the title, the description, the first sentence and the snippet. The snippet is constituted of the first 4 sentences comprised within the first 512 characters of the text - everything after was ignored, sentences were not truncated. We present here the set of features that have been computed for each of these fields, and for some cross field (title-description, title-first sentence, description-first sentences), all features are normalized.

Lexicographic measures: we consider two sets of representation: the set of words, and the set of multi-word expressions (MWE). The MWE were computed by splitting the text around stopwords and punctuation marks, essentially similar to RAKE keyword extraction (Piskorski et al., 2021). The features include the proportion of matching words and proportion of matching MWE between both articles. In the case of MWE, a margin of error was allowed in that two non strictly equal MWE were considered equal if the longest common subsequence between them was of 75% the length of the longest one. For both words and MWE we compute the raw count of elements in common and the length of the corresponding span of text.

Embedding-based measures: all embedding-based measures are based on LASER embeddings (Artetxe and Schwenk, 2019) which are aligned multilingual embeddings covering over 100

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 1178 - 1183

languages and that perform well for multilingual semantic comparison. They are BiLSTM based, and as such are much faster than transformer-based solutions, it also has to be noted that out of the box BERT solutions do not perform well on the task of semantic similarity, and LASER embeddings have a better performance (Reimers and Gurevych, 2020). On top of the pairwise comparison between the aforementioned fields, we also compute the equivalent of Word Mover Distance (Zhao et al., 2019) between the first 4 sentences of the articles, by binning sentences in pairs of decreasing similarity.

Non linguistic features: the only non linguistic feature considered is the difference in publication date of the two articles.

When developing the system, the results were checked for near misses, by considering all the clusters produced, and checking for missing links between elements of the clusters. This approach was able to catch such misses, including in the training data itself, but it was not used in the final system due to lack of time.

3 Data

The training data provided by the organizers is a list of 4964 article pairs covering 9431 articles in 16 languages, while the 7 main languages represent 99% of the dataset. The pairs are associated with several similarity scores; one of them giving the overall similarity is ranked between 1.0 (maximum) and 4.0 (minimum) and is the only one considered in our approach. Test data contains 4890 pairs over 9715 articles in 10 languages. There is a notable difference between both datasets, in that train data contained almost no pairs of article in different language, while such pairs represented 15% of the evaluation dataset. A major difference was also the introduction of Chinese in the test set.

3.1 Acquisition

The data was given as a list of urls to download, and this proved to be a daunting exercise fraught with difficulties and eventually taking more time and dedication than the development of the system itself. The scrapper provided by the organizer was not used, we relied on the *trafillatura* (Barbaresi, 2020) library, which was desirable thanks to its good metadata extraction capacities and overall performances. A total of 3 different scrapping approaches were used: first trying to scrap directly from the original url, then trying to scrap from the internet archive, finally, in case the url was reachable but the data was not correctly extracted, we wrote an ad-hoc scrapper whose extraction rules were manually written for each problematic news source. Despite these efforts, about 7% of the test data was impossible to download. We report these numbers in Table 1. Most of the articles that had to go through the ad-hoc scrapper were from Chinese sources, for several of these it was possible to extract only the title, this created a difference in the features available with several missing values.

set	direct	arch.	ad-hoc	total	absent
train	8108	951	159	9431	274
test	7348	847	838	9715	682

Table 1: Number of additional downloaded pairs of urls by scrapping method, total and missing pairs

3.2 Preprocessing

The data underwent significant prepossessing. The title was cleaned of mentions to the news source. In order to do that, if the source name was extracted from the metadata and was present in the title it was removed, if a token was ending with an internet top level domain, the token was removed, if a pipe bar was present, the leftmost part of the bar was kept. If it was detected that a source always had the same title, indicating that the scrapper was not correctly extracting it, then the title was replaced by the first sentence of the description if available. If the description was not available, the first sentence of the article was used instead, in case the article had no text but had a description, the description was used for the text, and in case there was not description but a title, the title was used for the description. This procedure aimed at minimising the impact of missing values.

The text was preprocessed by removing likely source name, author name and date mention at the beginning of the text, using an heuristic rule-based procedure focusing on the presence of numbers, uppercase letter, short sentences and typical start of text markers, such a double dashes. Out of the remaining text, a snippet was extracted, taking the sentences spanned by the first 512 characters, with a maximum of 4 sentences. Unicode letters were normalized to canonical forms. The language annotation provided in the dataset was not trusted, and we instead relied on the one provided by the lan-

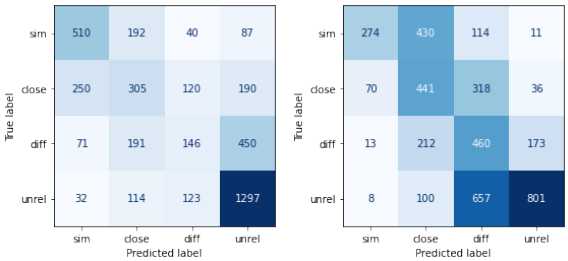

Figure 1: Confusion matrix for random forest classification (left) and rounded random forest regression (right)

guage detector of `fasttext` (Grave et al., 2018). Chinese texts were segmented using the `jieba` library in order to insert spaces between words so as to further deal with them in a language agnostic way.

4 Experimental Results

4.1 Experiments on train dataset

We settled on using random forests on the classification task as our contribution model after preliminary trial showed that random forests had better performances than neural networks and that the label distribution produced by random forests on the classification task was much more coherent with the ground truth than the one produced by random forests applied to the regression task. Given that we will tackle the problem as a classification problem, we will consider 4 classes labelled from 1 to 4, denoting the following relations in article pairs: similarity (sim), close similarity (close), different (diff) and unrelated (unrel). Unless specified otherwise, all the values computed over the train set are an average over 5-fold cross validation.

In Figure 1 we report the confusion matrix over the 4 classes using all the features with a random forest classifier (RF-C) and regressor whose output have been rounded to the closest integer (RF-R). RF-R has more errors than RF-C, but these are less significant as classes are mistaken mostly between related classes: for instance the classes sim and close are more often confused than with RF-C, but the classes sim and unrel are significantly less confused.

In Figure 3 we report the label distribution of the classifier and the regressor when trained and tested over the full training set and compare it with the ground truth. In this same setting, in Table 2 we report the Jensen-Shannon divergence of these label distributions, measuring the distance with the train data label distribution (Fuglede and Topsoe, 2004), we also report the micro F_1 and macro F_1 performance of RF-C and RF-R. Despite a seeming overfitting of RF-C, the distribution of labels produced by RF-C is clearly closer to the one of the ground truth than the one of RC-R. For all these different reasons we decided to tackle this task as a classification problem instead of a regression one.

Figure 2: Caption

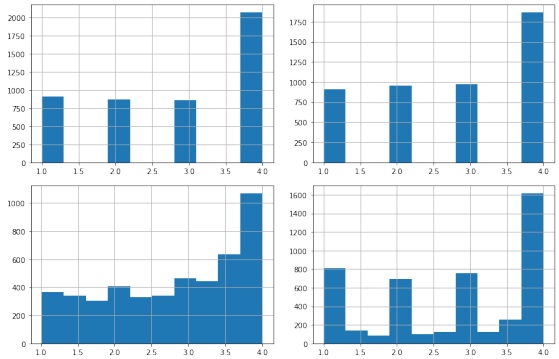

Figure 3: Label distribution for random forest classification (up. left), random forest regression (low. left), ground truth (low. right) and rounded ground truth (up. right)

measure	RF-C	RF-R
micro F1	93.6	80.2
macro F1	93.7	78.9
JS div.	0.0016	0.0098

Table 2: Performances on the train data, and distance of the label distribution with respect to the train dataset distribution

In Table 3 we report the performance of our classifier (RF-C) over different subsets of the training data, in function of the language pairs considered: all (MULTI), only same languages (SAME), only English (EN) and only different languages (CROSS). Both micro an macro F_1 are the highest for EN, nevertheless MULTI is the second best performing subset in terms of macro, with only one point less, while the micro difference of 6 points is important. The performance of SAME an MULTI are similar, this seems to indicate that specifically for English the performances are better. This could be due to the fact that English is the less flexional language of all the languages in the training set, as such, string matching is more likely to report the correct answer, both for exact and approximate matching. CROSS has the second highest micro score, but has the lowest macro score. This approach tends to rely more heavily on the embed-

subset	micro F_1	macro F_1
MULTI	55.2	47.5
SAME	54.1	47.2
EN	61.1	48.3
CROSS	58.0	40.1

Table 3: multilabel evaluation of RF-C: micro F_1 and macro F_1 performance of different subsets of language pairs of the train dataset

subset	F_1 (1 vs rest)	F_1 (1+2 vs rest)
MULTI	60.1	78.0
SAME	59.1	74.1
EN	64.5	75.1
CROSS	56.3	76.1

Table 4: binary evaluation of RF-C: the F_1 measure is reported for two binary classifier: class 1 (similar articles) vs rest and class 1+2 (similar and close articles) vs rest.

features	micro F_1	macro F_1
ALL	55.2	47.5
LEX+EMB	53.3	44.8
LEX	49.6	40.4
EMB	51.3	42.6
DATE	37.5	22.6

Table 5: evaluation of different subsets measured as the micro and macro F_1 of RF-C in a multilabel setting

subset	support	micro F_1	macro F_1
all eval dataset	4902	44.1	40.5
all downloaded	4455	46.5	41.5
all with text	3946	47.6	42.4

Table 6: micro and macro F_1 performance of our approach on the evaluation dataset, for different subsets of articles pairs

ding feature as lexical matches are unlikely except for named entities.

In Table 4 we report the measures for the same subsets, but applied when considering binary classifications problems. We consider two such classifiers: when considering class 1 (similar articles) versus the rest of the classes and when considering class 1 and 2 (similar and close articles) versus the rest of the classes. The performance of the binary classifiers are clearly superior to the performance of the multilabel classifier: on MULTI the former has a micro F_1 performance about 5 points better, and the later has a better performance by 22 points.

Study of the correlation matrix of the features with the ground truth, not reported in this paper for readability reasons, shows that all lexicographic-based features are the highest correlated features, with a correlation percent of about 50%. Among the several embedding-based features, only the distance between titles correlated highly with ground truth with 46%, second to it only similarity between title and description has a significant correlation of around 32%.

In Table 5 we report the performance of our classifier on the train dataset for different subsets of the features: all the features (ALL), only lexicographic-based ones (LEX), only embedding-based one (EMB), a combination of both (LEX+EMB) and date (DATE). ALL has clearly the best performance both in terms of macro and micro F_1, despite integrating the date feature, which on itself performs quite poorly. However, we believe that the impact of this feature is heavily biased by the way both the train and test datasets have been constructed, as the organizer of the shared task have fetched news over a time period of several years, while about half the items are related, and about a quarter are similar. As a consequence, it happens that close dates are related to similar news more than it would appear in an uniform sample over the same time period. For that reason, we don't expect that this feature would generalise well, but we left it nevertheless as we expect the train and test distributions to be similar.

4.2 Experiments on test dataset

When evaluating on the test dataset, we use a classifier (RF-C) trained on the full training dataset. Because of the difficulties in downloading the test data, we report the performance of the classifier on different subsets of the downloaded data.

In Table 6 we report the micro and macro F_1 performances of our classifier, as well as the support, counted in number or article pairs. We consider three subsets: all evaluation data (including also article pairs whose articles were not downloaded, and for which a default score of 2.69 was used - which was the average predicted value of our classifier for the other articles), all pairs whose corresponding articles have been successfully downloaded, all pairs for which the text of the corresponding articles have been successfully downloaded. The best performance on the test data is on average 5 points lower both in term of micro an macro than on the

subset	micro F_1	macro F_1	F_1
MULTI	46.5	41.5	74.0
SAME	43.7	42.0	78.6
EN	58.4	42.7	79.0
CROSS	43.8	38.3	71.7

Table 7: performance on the evaluation dataset of the classifier, using micro and macro F_1 as well as F_1 for the corresponding binary classification problem

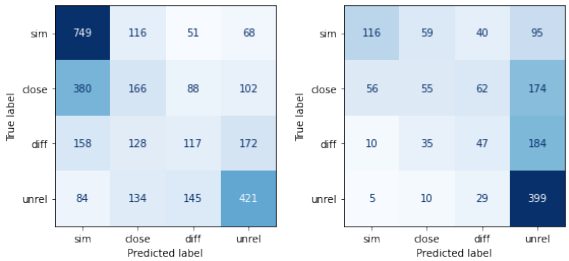

Figure 4: Confusion matrix of RF-C on the test set, for same-language pairs (left) and cross-language pairs (right)

train dataset. Expectedly, the model evaluated on the full dataset performs the worst, but reasonably so with only 2 points down the performance on the subset of successfully downloaded articles while 9% of the data rely on the default value. Further restricting the evaluation on the articles that had a successfully downloaded text content only provides one additional point of performance.

In Table 7 we report the performance of the classifier on the test dataset for different subsets of language pairs, as already previously described. The measures are micro and macro F_1, as well as the F_1 of the class 1+2 when considering tackling the problem as binary classification. Interestingly, while the mutlilabel performances are lower than on the training set by about 9 points, the binary classification performance is only lower by 5 points in a cross language setting and actually higher by 4 points in a same language setting.

In Figure 5 we report the true positive rate by language pairs for pairs of languages having more than 10 data points, evaluated on the subset of the dataset for which articles were successfully downloaded. We consider a true positive in case the predicted and ground truth had exactly the same label, as such it is not possible to assess how close to the actual label the predicted values are. Therefore, this figure only allows to give an overall picture of the respective performances for pairs of languages. When considering pairs of articles in the same languages, French, Spanish and English performs the best. Among these, Chinese has the worst score, this could be related to the fact that a few Chinese sources representing a significant share of articles were impossible to be correctly downloaded (text and other metadata are absent). Surprisingly some pairs of different languages perform better than for these languages considered individually, this is the case for German and Chinese. Nevertheless, articles pairs in different languages tend to perform worse than pairs in the same language.

In Figure 4 we report the confusion matrix of our classifier, by considering separately two cases: the confusion matrix for same-language pairs and for cross-language pairs. From this figure it is clear that our classifier tends to over-estimate the similarity of articles written in the same language, while underestimating the similarity of articles written in different languages.

5 Discussion

Despite the problem of the shared task being posed as a regression one, we have chosen to address it as a classification problem. This has two direct negative consequences on the performance of the model: firstly due the mandatory discretization step of the real-valued evaluation score provided in the training data and expected by the evaluation system; secondly due to the fact that the notion of related classes, such as "similar" and "close", are lost and penalized as much as if "similar" had been predicted as "unrelated". This is clearly shown by the fact that the multilabel classifier trained on 4 labels performs significantly worse than a binary classifier trained on two classes, which has actually acceptable performances. Despite being language agnostic our approach performs better in a same language setting than in a cross language one. This indicates either the high importance of the lexicographic as being a good predictor, or that multilingual embeddings perform significantly differently on different language pairs.

The lexicographic-based features perform surprisingly well, only two points under the performance of the embeddings-based features when evaluated on the full training dataset contain. A potential reason for that could be that named entities can differ only slightly between languages and that the soft lexicographic measure used is good at capturing these variations. The performance of our

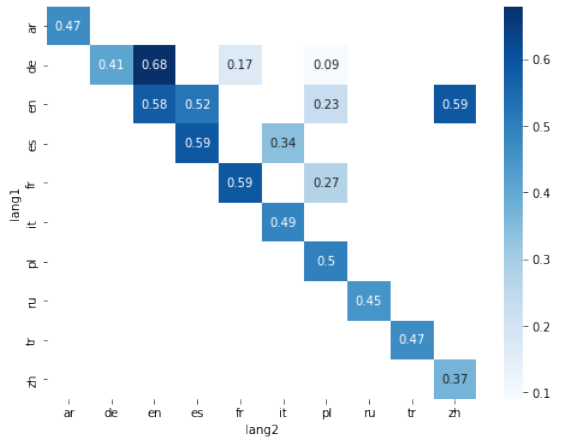

Figure 5: True Positive Rate by language pair of the multilabel classifier on the test dataset

approach is lower on the test dataset, such a drop in performance could be due to different reasons among others: the fuzzy lexicographic matching having lower performance in a cross language setting (much more prevalent in the test dataset than in the training); the difficulty to download data and correctly extract it, making the data incomplete; the heuristic reconstitution of missing description and text content could have introduced noise in some cases; maybe the fact that only tiny snippets of the full article were considered; or potentially a more fundamental limitation of the approach. Our approach could be improved by considering the actual language pairs as features, using more advanced features based on named entity extraction, and using exclusively data without missing values. However, we lacked time to investigates all of these configurations. Interestingly, of the embedding-based features, the title and the description are the most important features, and work better than more advanced ones, such as word mover distance applied to the sentences of the snippets.

Given these results, we can argue that the language agnostic approach we developed is an interesting solution for a coarse grain similarity evaluation, but not for a fine grained one. Given the fast computation time, a few minutes to compute all the features on a CPU machine without using multi-threading, our approach could be used as a preprocessing step before using more precise but also more time consuming approaches. Particularly, this approach is interesting when a news processing system has to process hundreds of thousands articles a day, preventing the use of costly solutions.

6 Conclusion

In this paper we presented the system we used to tackle the multilingual news clustering shared task at SemEval-22. Our approach is language-agnostic and inherently multilingual. It relies on a set of relatively simple lexicographic- and embedding-based features, and as such is able to process documents efficiently. We tackle the task as a classification problem rather than a regression problem. Our approach performs satisfyingly when evaluated over a simpler binary classification problem.

References

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Adrien Barbaresi. 2020. htmldate: A Python package to extract publication dates from web pages. *Journal of Open Source Software*, 5(51):2439.

Xi Chen, Ali Zeynali, Chico Q. Camargo, Fabian Flock, Devin Gaffney, Przemyslaw A. Grabowicz, Scott A. Hale, David Jurgens, and Mattia Samory. 2022. SemEval-2022 Task 8: Multilingual news article similarity. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Bent Fuglede and Flemming Topsoe. 2004. Jensen-shannon divergence and hilbert space embedding. In *International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings.*, page 31. IEEE.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. *arXiv preprint arXiv:1802.06893*.

Jakub Piskorski, Nicolas Stefanovitch, Guillaume Jacquet, and Aldo Podavini. 2021. Exploring linguistically-lightweight keyword extraction techniques for indexing news articles in a multilingual set-up. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pages 35–44.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. *arXiv preprint arXiv:2004.09813*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. *arXiv preprint arXiv:1909.02622*.