Cross-lingual categorization of news articles

Blaž Novak
Jožef Stefan Institute
Jamova 39
Ljubljana, Slovenia
+386 1 477 3778
blaz.novak@ijs.si

ABSTRACT
In this paper we describe the experiments and their results performed with the purpose of creating a model for automatic categorization of news articles into the IPTC taxonomy. We show that cross-lingual categorization is possible using no training data from the target language. We find that both logistic regression and support vector machines are good candidate models, while random forests do not perform acceptably. Furthermore, we show that using Wikipedia-derived annotations provides more information about the target class than using generic word features.

General Terms
Algorithms, Experimentation

Keywords
News, articles, categorization, IPTC, Wikifier, SVM, Logistic regression, Random forests.

1. INTRODUCTION
The JSI Newsfeed [1] system ingests and processes approximately 350.000 news articles published daily around the world, in over 100 languages. The articles are automatically cleaned up and semantically annotated, and finally stored and made available for downstream consumers.

One of the annotation tasks that we would like to perform in the future is to automatically categorize articles into the IPTC "Media Topics" subject taxonomy [2]. IPTC – the International Press Telecommunications Council – provides a standardized taxonomy of roughly 1100 terms, arranged into a 5 level taxonomy, describing subject matters relating to daily news. The vocabulary is accessible in a machine readable format – RDF/XML and RDF/Turtle – at http://cv.iptc.org/newscodes/mediatopic.

There are two relations linking concepts in the vocabulary – the 'broader concept' taxonomical relation, and a 'related concept' sibling relation. The 'related concept' links concepts both to other concepts from the same taxonomy, and directly to external Wikidata [3] entities.

The purpose of this work is to evaluate multiple machine learning algorithms and multiple sets of features with which we could automatically perform the categorization. As we would like to categorize articles in all the languages the Newsfeed system supports, but we only have example articles in English and French, the method needs to be language independent.

2. EXPERIMENTAL SETUP
The dataset that we have access consists of 30364 English and 29440 French articles, each of which is tagged with 1 to 10 categories. We consider each document belonging to all categories that are explicitly stated, and all of their parents. We will compare the performance of model predictions on the same language and in the cross-lingual setting, where we train the model on the entire dataset available for one language, and measure its performance on the other language.

Basic features of the dataset can be seen in the following 2 figures. Figure 1 shows the distribution of number of articles in each category, and Figure 2 shows that most categories contain a roughly even number of articles in both languages, but there are some outliers. We ignored categories with less than 15 examples per language, which resulted in 308 categories.

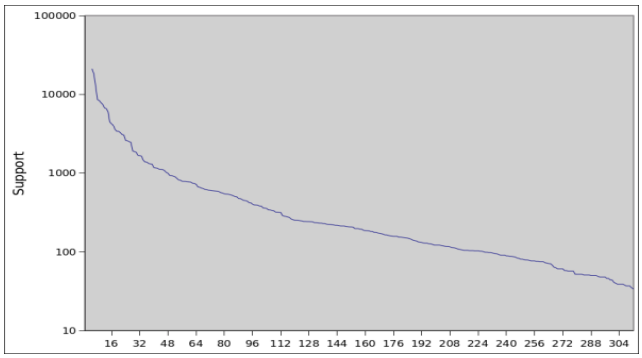

Figure 1. Number of articles in each category. Discrete categories on x axis are ordered by descending number of articles.

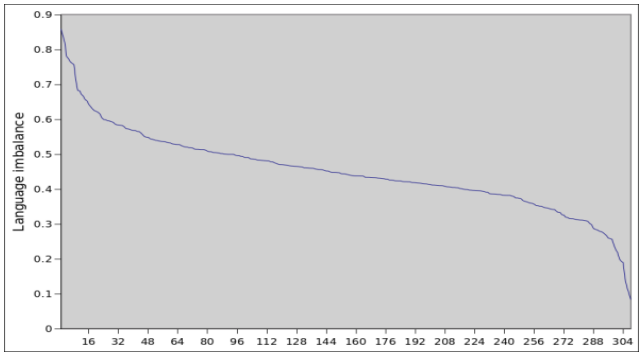

Figure 2. Language imbalance for each category. Discrete categories on x axis are ordered from "mostly English" to "mostly French".

We compare three different machine learning models – random forests, logistic regression (LR), and Support Vector Machines (SVM).

We try two different types of features, and their combinations.

The first kind of a feature set we use is a projection of the bag-of-words representation of the document text into a 500 dimensional vector space. The KCCA [4] method uses an aligned multi-lingual corpus to find such a mapping, that words with similar meanings map to a similar vector, regardless of their language. We represent a document as a sum of all word vectors.

The second set of features we use is the output of the JSI Wikifier [5] system. The Wikifier links each word in a document to a set of Wikipedia pages that might represent the meaning of that word. For each such annotation, we also get a confidence weight.

We consider these annotations as a classical vector space model -- as a bag-of-entities. We use two versions of the TF-IDF [7] scheme: in the first case, we use the number of times an entity annotation is present for any word in a document as the TF (term frequency) factor, and in the second version, we use the sum of annotation weights of an entity across the document. In both cases, we perform L1 normalization of the vector containing TF terms. For IDF terms, we use $\log\left(1+\frac{N}{n}\right)$ where N is the number of all documents and n the number of documents where an annotation was present at least once.

Finally, we use a combination of both KCCA-derived and Wikifier-derived features as the last feature set option.

For model training, we use Pythons scikit-learn [6] software package. In the case of logistic regression, we use L2 penalty, with automatic decision threshold fitting, using the liblinear library backend.

For the SVM model, we use a stochastic gradient descent optimizer. We performed a grid search for the optimal regularization constant C, but since there were no significant accuracy changes, we used the default of 1.0 in all other experiments.

For the random forest model, we used 4 different parameter combinations:

- default – 10 trees, splitting until only one class is in the leaf
- 30 trees, maximum tree depth of 10
- 50 trees, maximum tree depth of 10
- 30 trees, maximu tree depth of 20

In all cases, GINI index was used as the node splitting criterion.

Since the majority of categories only have a small number of documents, we automatically weighed training examples by the inverse of their class frequency. We also performed some experiments without this weighting scheme, but got useless models in all cases except for the couple largest categories.

All reported results are the average of a 3-fold cross-validation.

So far, we only created one-versus-all models for each category independently, and only used the taxonomy information of categories to select all examples from sub-categories when training the more general category.

3. RESULTS

Table 1 shows ROC scores for cross-validation of all three models on four sets of feature combinations, for English and French separately. SVM and logistic regression are comparable in behavior and promising, while the random forest model performs significantly worse. "Wiki-W" denotes the weighted version of Wikifier annotations, and "Wiki-K" the combination of KCCA-derived features and Wikifier annotations. Every second line in the table is the standard deviation of the result when averaged across all categories.

Table 1. ROC scores by model and feature type, cross-validation

	Rand. Forest		Log. Reg.		SVM	
	EN	FR	EN	FR	EN	FR
KCCA	0.75	0.71	0.96	0.95	0.95	0.94
(stdev)	0.11	0.11	0.04	0.04	0.05	0.04
Wiki	0.70	0.70	0.95	0.95	0.94	0.94
(stdev)	0.12	0.12	0.04	0.04	0.05	0.04
Wiki-W	0.71	0.71	0.95	0.95	0.94	0.94
(stdev)	0.12	0.11	0.04	0.04	0.05	0.04
Wiki+K	0.71	0.69	**0.97**	**0.96**	0.96	0.95
(stdev)	0.12	0.11	0.03	0.03	0.03	0.04

Looking at the feature selections, we see almost no significant difference -- both kinds of features -- KCCA and Wikipedia annotations have useful predictive value. The combination of both feature types slightly improves the ROC score.

Table 2 shows F1 cross-validation scores of all three models. Logistic regression scores much higher than SVM here, possibly indicating that the SVM model would benefit from a post-processing step of optimizing the decision threshold on a separate training set.

Table 2. F1 scores by model and feature type, cross-validation

	Rand. Forest		Log. Reg.		SVM	
	EN	FR	EN	FR	EN	FR
KCCA	0.16	0.12	0.30	0.25	0.20	0.18
(stdev)	0.21	0.18	0.21	0.20	0.21	0.19
Wiki	0.07	0.07	0.41	0.44	0.25	0.29
(stdev)	0.15	0.15	0.21	0.21	0.22	0.22
Wiki-W	0.08	0.08	0.40	0.43	0.24	0.28
(stdev)	0.17	0.17	0.21	0.21	0.21	0.22
Wiki+K	0.09	0.07	**0.44**	**0.46**	0.27	0.30
(stdev)	0.16	0.15	0.21	0.21	0.22	0.22

The combination of both feature sets performs significantly better than either alone, with generic word-based features providing the least amount of information.

The feature usefulness changes when looking at cross-lingual classification performance. Table 3 shows the ROC score for all three models, when the model trained on English is used to predict categories of French articles, and vice versa. Decision trees give essentially a random result, and SVM scores somewhat higher than logistic regression.

Table 3. ROC scores - cross-lingual classification

	Rand. Forest		Log. Reg.		SVM	
	EN	FR	EN	FR	EN	FR

KCCA	0.50	0.50	0.50	0.50	0.50	0.51
(stdev)	0.00	0.00	0.01	0.03	0.04	0.08
Wiki	0.51	0.51	0.76	0.80	0.81	0.84
(stdev)	0.04	0.04	0.12	0.11	0.11	0.10
Wiki-W	0.51	0.52	0.78	0.82	**0.82**	**0.84**
(stdev)	0.04	0.05	0.11	0.10	0.10	0.10
Wiki+K	0.50	0.50	0.57	0.70	0.66	0.81
(stdev)	0.01	0.01	0.10	0.13	0.14	0.12

The biggest change here is the influence of KCCA cross-lingual word embedding: by itself it provides no informative value, as indicated by ROC value of 0.5 in all cases, and it even reduces the performance of the combined Wikifier + KCCA model.

In the Table 4, F1 scores from the same experiment are shown. Logistic regression still has a big advantage over SVM, as in the same-language categorization setting. The change from previous experiments is the influence of weighting of Wikipedia features -- it increases the performance of all models.

Table 4. F1 scores - cross-lingual classification

	Rand. Forest		Log. Reg.		SVM	
	EN	FR	EN	FR	EN	FR
KCCA	0.00	0.00	0.00	0.01	0.00	0.02
	0.02	0.02	0.02	0.06	0.01	0.06
Wiki	0.03	0.04	0.48	0.44	0.30	0.26
	0.10	0.11	0.21	0.20	0.22	0.22
Wiki-W	0.03	0.05	**0.49**	**0.44**	0.29	0.26
	0.11	0.13	0.20	0.21	0.22	0.22
Wiki+K	0.00	0.00	0.18	0.40	0.20	0.23
	0.04	0.04	0.22	0.22	0.19	0.21

An interesting observation is that the performance of the cross-lingual model is occasionally higher than that of the baseline cross-validation experiment. This anomaly however disappears for categories with large amount of positive training examples. It also disappears if we reduce the amount of training examples in the cross-lingual experiment by 1/3 – the effect seems to be caused by cross-validation reducing the training dataset size.

KCCA cross-lingual word embedding feature generation used here was tested in other experiments and systems and gives a useful feature set for comparison of documents across languages, so its negative impact on the performance of these models needs to be investigated in the future.

As the weighted Wikipedia feature set appears to be the best for the stated goal of cross-lingual article categorization, the results of next experiments are shown only for it, but we performed the same experiments on all other combinations, and the results broadly follow the conclusions from the previous section.

The following figures show correlation of testing and cross-lingual performance of logistic regression and SVM models. Both F1 score and area under ROC curve are shown for each of 308 categories in the experiment, since they provide complementary information. As the figures show, there is a good agreement between the cross-validation and the cross-lingual classification performance, giving us an ability to estimate cross-lingual performance based on the cross-validation score in the production environment. The difference between distributions for French and English language models is consistent with the class imbalance for each of the categories.

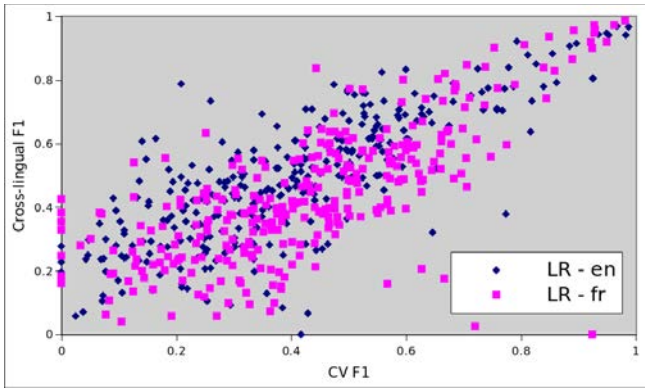

Figure 3. F1 score correlation for logistic regression

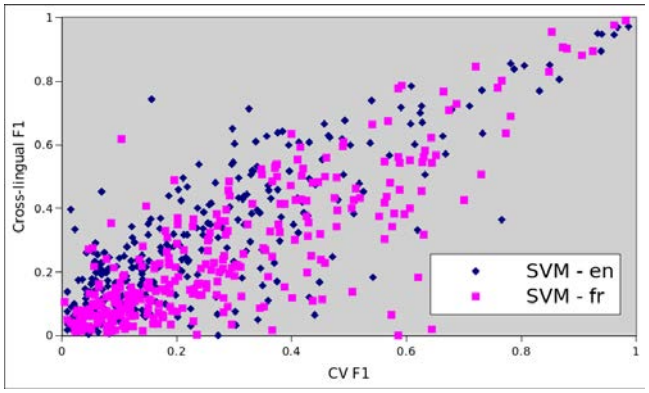

Figure 4. F1 score correlation for SVM

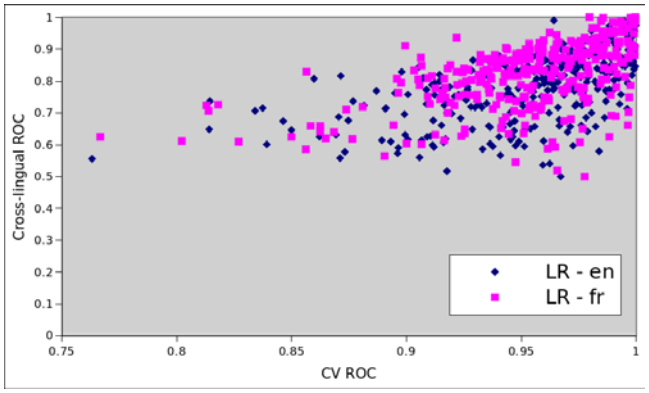

Figure 5. ROC score correlation for logistic regression

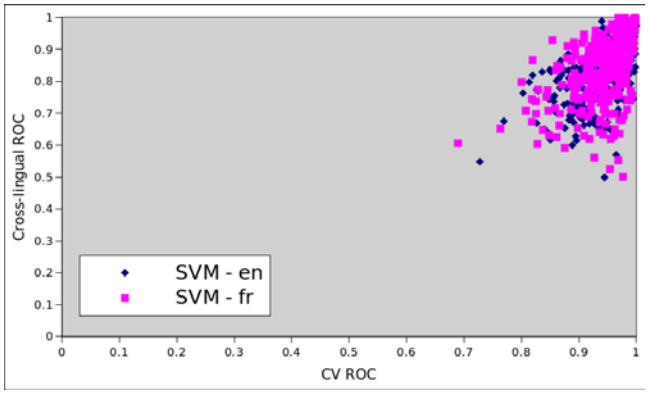

Figure 6. ROC score correlation for SVM

The SVM model seems to have a more consistent behavior, so we will use it in the final application instead of logistic regression.

Figures 7 through 10 show the F1 and ROC score behavior of logistic regression and SVM models for cross-validation and cross-lingual classification with regard to the number of positive examples in the category, separately for English and French language. While the SVM model underperforms on the F1 metric on average, it produces a better ranking of documents with respect to a category, as seen on ROC plots, especially for smaller categories. This further indicates the need for decision threshold tuning in the SVM model before we use its predictions.

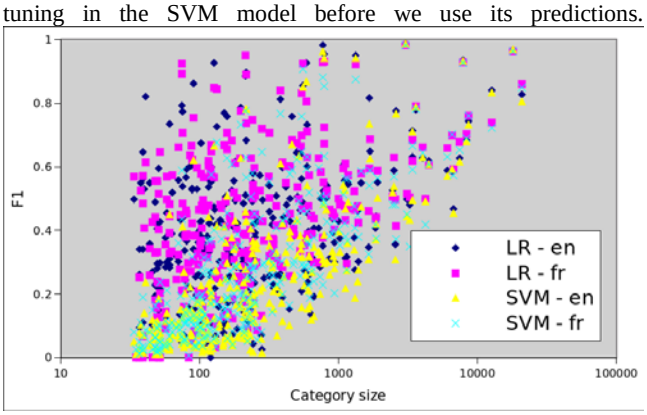

Figure 7. F1 score with respect to category size, cross-validation

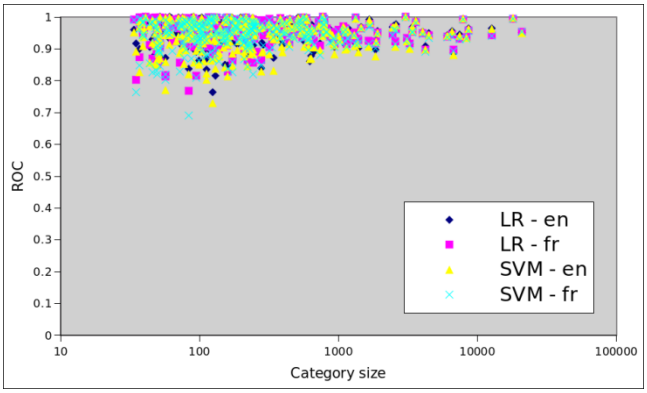

Figure 8. ROC score with respect to category size, cross-validation

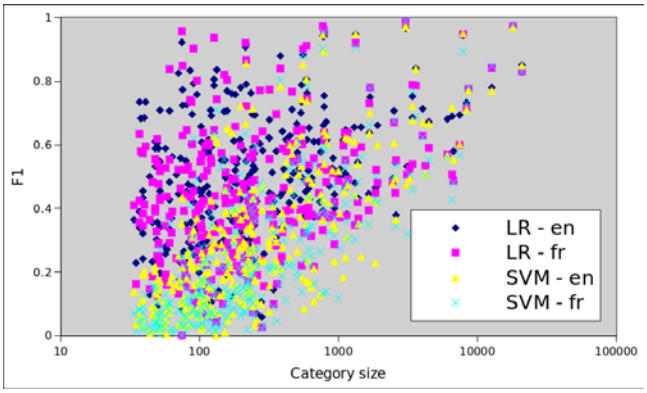

Figure 9. F1 score with respect to category size, cross-lingual prediction

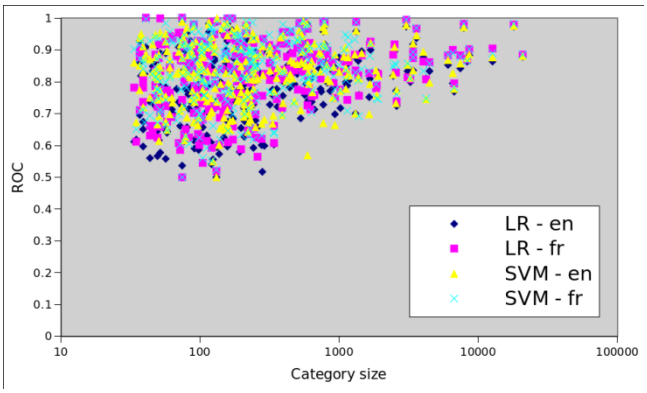

Figure 10 ROC score with respect to category size, cross-lingual prediction

As expected, classification performance of all models improves with the number of training examples, but in cases of small categories, it appears that some are much easier to learn than others.

4. CONCLUSIONS AND FUTURE WORK

We found that using a logistic regression model with weighted Wikifier annotations gives us a good enough result to use IPTC category tags as inputs for further machine processing in the Newsfeed pipeline. Before we can use this categorization for human consumption, we need to investigate automatic tuning of SVM decision thresholds on this problem, and add an additional filtering layer that takes into consideration interactions between categories beyond the sub/super-class relation. Additionally, the negative effect of KCCA-derived features for cross-lingual annotation needs to be examined.

5. ACKNOWLEDGEMENTS

This work was supported by the Slovenian Research Agency as well as the euBusinessGraph (ICT-732003-IA) and EW-Shopp (ICT-732590-IA) projects.

6. REFERENCES

[1] Trampuš M., Novak B., “The Internals Of An Aggregated Web News Feed” Proceedings of 15th Multiconference on Information Society 2012 (IS-2012).

[2] https://iptc.org/standards/media-topics/

[3] https://www.wikidata.org/wiki/Wikidata:Main_Page

[4] Rupnik, J., Muhič, A., Škraba, P. “Cross-lingual document retrieval through hub languages”. NIPS 2012, Neural Information Processing Systems Worshop, 2012

[5] Brank J., Leban G. and Grobelnik M. “Semantic Annotation of Documents Based on Wikipedia Concepts”. Informatica, 42(1): 2018.

[6] Pedregosa, F., Varoquaux, G., Gramfort, A. et al. “Scikit-learn: Machine Learning in Python”. Journal of Machine Learning Research, 12. 2011, pp. 2825-2830.

[7] K. Sparck Jones. "A statistical interpretation of term specificity and its application in retrieval". Journal of Documentation, 28 (1). 1972