Evaluating Categorisation in Real Life – an argument against simple but impractical metrics

Vide Karlsson[1,2], Jussi Karlgren[1,2], and Pawel Herman[2]

[1] Gavagai, Stockholm
[2] KTH, Stockholm

Abstract. Text categorisation in commercial application poses several limiting constraints on the technology solutions to be employed. This paper describes how a method with some potential improvements is evaluated for practical purposes and argues for a richer and more expressive evaluation procedure. In this paper one such method is exemplified by a precision-recall matrix which sacrifices convenience for expressiveness.

1 The use case: Practical cold start text categorisation

Text categorisation in commercial application often involves texts of very various quality and genre, in large quantities, with categories of differing size and urgency, and which sometimes overlap. Text categorisation is used primarily to lessen the human effort involved in keeping track of and reading text streams, and thus will typically proceed without continuous human oversight and intervention, frequently with the results of the categorisation filed and archived for potential future reference, never to be examined by human readers: the end result of the effort are more typically statistics or summaries of the processed material.

In this present study, which is a pre-study for a practical text categorisation methodology with minimal start-up time, the stakeholder party wishes to offer its customers a service which can categorise texts without recourse to manually labelled training data, using only the category labels for categorisation and with the following minimal technology requirements:

Requirement 1: The system cannot rely on a previously manually labeled training set. Providing manually labeled training data requires too much work and is too data-intensive: a human takes about 1 minute per abstract to categorise 100 abstracts.[5, 8] Instead categories will be defined by a small set of manually chosen *seed labels*.

Requirement 2: The coverage and precision for the categories to be targeted should assessable automatically. The quality of the system output must be monitored and the customer will want to know how it performs on their categories.

Requirement 3: The baseline system uses simple string search for the category label. A marketed system must perform better than the baseline.

Requirement 4: The algorithm should be language independent. The stakeholder's customers operate in many of the languages of the world and not all of them have online terminologies or ontologies available.
Requirement 5: The system should be able to handle any topical category on any level of abstraction, as long as there are the category seed label is found in the data.
Requirement 6: The system should not require the topics to be structured hierarchically or to be disjoint in the data.
Requirement 7: The system should accommodate new categories being entered into the palette with little or no burn-in period and little cost.

2 Keyword-based categorisation

To meet the requirements of the use case, keyword-based categorisation was identified as a low-footprint technology, possibly enhanced by human intervention. Keyword-based categorisation schemes start from the original category label and then enrich it from knowledge of human language in general or from inspection of material from the data set under consideration. Many such approaches have been suggested. [1–4, 6] Besides conforming to the above requirements by providing a natural starting point with the category label as a seed term, a keyword-based approach has the additional advantage of the representation being handily inspectable and editable by a human editor without special training, and without reliance on hidden variables which even a human knowledge engineer would not be able to inspect, improve, and maintain.
This appears to be quite promising for the purpose of commercially viable application to customer tasks, but on closer inspection, reveals that the most of the methods under consideration rely on some hand-edited resource such as Wikipedia or Wordnet, which render them unsuitable with respect to the stakeholder requirements given above. Many of the methods take purchase in assumption of disjoint categories, using term distribution over categories as a criterion for selection. This does not conform to the expectations of the stakeholder and any such method will have to be left aside.
The test conditions in our first set of experiments were

grep Simple string matching to original one-word category label;

dice Category label enriched with terms selected and ranked by their Dice score, a simplified pointwise mutual information metric, calculated by collocation statistics of each term to other terms in the categorised gold standard [4]:

$$Dice(w_a, w_b) = \frac{D(w_a, w_b)}{D(w_a) + D(w_b)} \tag{1}$$

where $D(w_i, w_j)$ is the number of documents that contain both terms w_i and w_j and where $D(w_i)$ is the number of documents that contain w_i.

rich category label enriched with *manually selected* terms from the list of terms with highest Dice scores and subsequent addition of *manually approved* (1) terms given by consulting an online lexicon with

synonymous terms, [7] and (2) morphological variants of each term in the representation.

The evaluation challenge is to determine whether the additional effort in the **dice** and **rich** conditions would improve results enough over the **grep** condition to warrant the investment in implementation, execution, and human editorial effort for practical projects.

Established public data sets for evaluating categorisation mostly share the qualities of being balanced in size, disjoint and non-overlapping, and – frequently – homogenous with respect to genre and style. The need for more realistic and messy data sets has been established and to a large extent redressed with the data set provided by Liebeskind et al [4], which consists of 2 000 user-generated movie reviews, manually categorised, but copied into a collection of 400 000 reviews from IMDB. For these experiments, we selected only categories with more than five manually assessed documents in the gold standard set, leaving 44 categories to be considered for our experimental evaluation.

The data set, the categories it is split up into, and their initial labels are input parameters for the evaluation of the categorisation method.

1. If the data set is unrealistic in any important respect this will affect the results. Examples, would be how an unrealistically balanced, cleaned, and homogenous data set is used to evaluate methods intended to be deployed on new text.
2. If the initial label of the category is misleading, too specific, or too general this will affect the results.
3. If the categories in the experimental set are impractical or unrealistic, or if there are complex dependencies or overlaps between categories this will affect the results.

In these experiments, the data set is designed by Liebeskind to be realistic. The experiment we performed (based on Liebeskind's method) is designed to meet the challenge of enriching original category labels. In terms of the category palette, we use the set given in the data set.

2.1 The F-score evaluation metric

Typically categorisation methods are optimised for performance by collapsing their recall and precision performance into one scalar by their harmonic mean, the F-score. This representation of the comparative performance of the methods shows us that the additional effort put into enriching the category labels indeed translates into higher evaluation scores. However, we have very little sense of what this means for practical purposes. Does this mean that the enriched labels give us better recall? Does this mean that the manual addition of items improve precision? In practice, the F-scores need to be immediately decomposed into their component recall and precision scores to be useful.

2.2 The R@P curve

For the above and related reasons Liebeskind et al use the *R@P curve* as their main evaluation tool. It illustrates the level of recall a classifier gives

at a given level of precision. This is conformant with transparent and predictable performance. Curves for the present experimental conditions are given in Figure 1 These curves show us that the improvements are at the low-recall high-precision end. This is useful information, and will help the client make a decision if the improvements to the baseline method are useful or not. We still cannot address the third evaluation question given above, however: we do not know what the difference across categories is.

2.3 P & R matrix — a more expressive representation

We propose the following rather more expressive (but correspondingly rather less handy) representation of evaluation results. For each level of precision and recall we note how many of the test categories achieve that performance or better and arrange the results in a precision-recall matrix, recording in each cell the number of categories on a given performance level. This means that the quality requirements from a client can be mapped to a cell in the matrix, and the corresponding performance of the categorisation method can be read out immediately.

The P & R-matrix is a much more challenging representation to digest and process at first glance. It is less useful for the immediate purpose of ranking systems with respect to each other, where a scalar metric is preferred for the convenience of the human evaluator. However, this, by contrast to the obfuscatory F-score, allows the evaluator in e.g. a procurement process to assess the number of categories which actually are of practical use in a live system.

The evaluation results for the three experimental conditions are given in Table 1. The cells in the bottom rows of the matrices hold all the 44 categories: every experimental category can — unsurprisingly — achieve 0.0 precision or better at any level of recall.

If we raise our expectations to require a 0.6 precision or better we find that less than half of categories can be brought to 0.6 precision or better, even at very low recall rates, and that the **dice** and **rich** methods increase the recall noticeably for that precision level. We also find that the manual enhancement reduces the number of categories with full precision at low recall levels. Further, we find that, conversely, only one category can give full recall at the reasonable precision level of 0.5 and that no experimental condition succeeds in improving this result. This means that the improvements to the baseline model mainly appear to improve recall at low precision levels. This is a clear result with effects on how a system based on the methods investigated can be expected to be marketed and deployed in real industrial use cases. That information would neither have been obtainable from F-score comparisons nor from R@P curves.

We also see, more discouragingly for our experimental results, that most categories cannot be brought up to this level. This, again, will be a very valuable result in practical application scenario. If a customer requests a system test for a number of categories of interest to them, that test should demonstrate (as given by the requirements above) how well the system can be expected to perform for the categories of interest, not averaged over them, but identified per category, identifying categories

such as those which never are able rise above 0.2 recall and 0.1 precision in the example, irrespective of method. That sort of informed feedback to (potential) customers is necessary if they are to trust the technology solutions they are interested in procuring for their business or other activities.

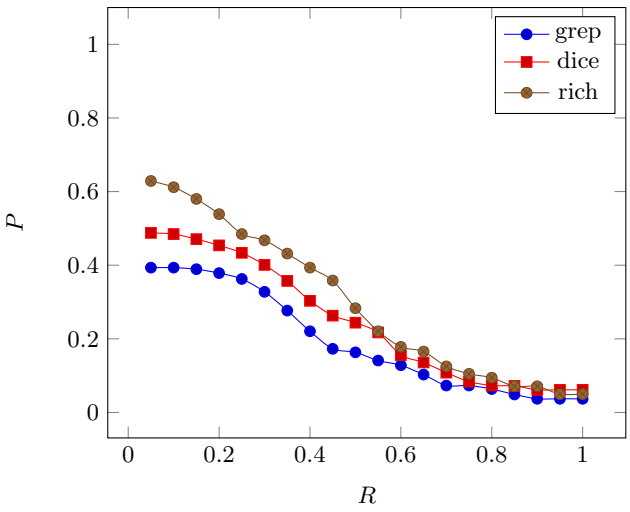

Fig. 1. R@P curves for the three experimental conditions

3 Conclusions

Keyword-based categorisation is motivated as a low-effort method for text categorisation. However, most methods rely heavily on precompiled resources, and are impracticable for the practical industrial use cases.
The evaluated methods show great volatility over topic categories. For more than 10% of the categories, it proved impossible to reach precision over 0.1 and for only a small portion of the categories a precision of more than 0.6 was attained. We note that not only – as has been discussed and shown in previous work, and to a great extent remedied by more realistic data sets – is the collection a parameter in an evaluation, but the category palette itself influences the result of the evaluation.
To determine the difference between the approaches chosen we need a more fine-grained evaluation method than the standard ones. The P & R evaluation matrix presented here is one such method which gives a more fine-grained result for evaluating and demonstrating the utility of a categorisation approach. A simpler evaluation method is a lossy compression of the information which is necessary to meet the requirements of a practicioner.

Table 1. P & R evaluation matrices for the three experimental conditions

F_1 **scores (micro averaged)**																																
	Method **grep**										Method **dice**										Method **rich**									
	0.313										0.354										0.399									
P & R Matrices																																
	Method **grep**										Method **dice**										Method **rich**									
	recall																															
p	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
1.0	6	4	2	2	1	0	0	0	0	0	12	4	2	1	1	0	0	0	0	0	9	5	3	2	1	1	1	0	0	0
0.9	6	4	2	2	1	0	0	0	0	0	12	4	2	1	1	0	0	0	0	0	10	6	3	2	1	1	1	0	0	0
0.8	8	7	3	2	1	0	0	0	0	0	13	4	2	1	1	1	0	0	0	0	11	7	5	4	2	2	2	0	0	0
0.7	10	7	4	2	1	1	1	1	0	0	15	8	5	2	2	2	1	0	0	0	14	10	7	5	4	3	2	0	0	0
0.6	17	9	7	7	1	1	1	1	0	0	18	11	6	4	3	2	2	2	0	0	20	13	10	9	6	5	3	3	0	0
0.5	21	13	9	8	5	3	2	2	1	1	26	19	13	10	7	4	3	2	1	1	32	21	14	14	7	7	5	3	1	1
0.4	23	14	12	8	7	6	3	2	1	1	29	21	17	13	11	7	6	4	1	1	36	27	21	19	13	10	6	4	2	2
0.3	26	17	15	14	11	9	4	2	1	1	32	23	21	17	14	11	8	6	3	2	38	31	26	23	19	14	10	5	3	2
0.2	32	21	20	16	14	12	6	6	2	1	36	26	22	19	15	12	12	8	4	3	38	36	29	25	23	17	12	9	5	3
0.1	33	22	20	17	14	12	8	7	2	1	36	27	23	19	17	15	15	12	5	3	40	38	35	32	28	25	20	16	7	3
0.0	44	44	44	44	44	44	44	44	44	44	44	44	44	44	44	44	44	44	44	44	44	44	44	44	44	44	44	44	44	44

References

1. Libby Barak, Ido Dagan, and Eyal Shnarch. Text categorization from category name via lexical reference. In *Proceedings of Human Language Technologies: Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 33–36. Association for Computational Linguistics, 2009.
2. Alfio Gliozzo, Carlo Strapparava, and Ido Dagan. Improving text categorization bootstrapping via unsupervised learning. *ACM Transactions on Speech and Language Processing (TSLP)*, 6(1):1, 2009.
3. Youngjoong Ko and Jungyun Seo. Text classification from unlabeled documents with bootstrapping and feature projection techniques. *Information Processing & Management*, 45(1):70–83, 2009.
4. Chaya Liebeskind, Lili Kotlerman, and Ido Dagan. Text categorization from category name in an industry-motivated scenario. *Language Resources and Evaluation*, 49(2):227–261, 2015.
5. Andrew McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. A machine learning approach to building domain-specific search engines. In *IJCAI*, volume 99, pages 662–667. Citeseer, 1999.
6. Qiang Qiu, Yang Zhang, Junping Zhu, and Wei Qu. Building a text classifier by a keyword and wikipedia knowledge. In *Advanced Data Mining and Applications*, pages 277–287. Springer, 2009.
7. Magnus Sahlgren, Amaru Cuba Gyllensten, Fredrik Espinoza, Ola Hamfors, Anders Holst, Jussi Karlgren, Fredrik Olsson, Per Persson, and Akshay Viswanathan. The Gavagai Living Lexicon. In *10th Language Resources and Evaluation Conference, Portoroz*, 2016.
8. Greg Schohn and David Cohn. Less is more: Active learning with support vector machines. In *ICML*, pages 839–846. Citeseer, 2000.