Assessing The Quality Of Opinion Retrieval Systems.

G. Amati[1], G. Amodeo[2], V. Capozio[3], C. Gaibisso[4], G. Gambosi[3]

[1]Ugo Bordoni Foundation, Rome, Italy
[2]Dept. of Computer Science, University of L'Aquila, L'Aquila, Italy
[3]Dept. of Mathematics, University of Rome "Tor Vergata", Rome, Italy
[4]IASI-CNR, Rome, Italy

The First International Workshop on Opinion Mining for Business Intelligence
August 31, 2010

Summary

Objectives of the work

- Topical Opinion Retrieval (TOR) is evaluated by classical IR evaluation measures, i.e. Mean Average Precision (*MAP*) or Precision at 10 (*P@10*).
- The effectiveness of the topical-only retrieval (effectiveness of the baseline) boosts the TOR performance.
- How can we assess the opinion-only classification accuracy (or precision, etc.)? How can we split the contribution of the opinion component from retrieval?

Methodological Framework

- We build artificial opinion-only classifiers from relevance and opinion data at different rates of opinion accuracy and precision.
- Then we study the effect on MAP of the TOR system with such classifiers
- We are able to assess the opinion-only component quality of a given TOR system by comparing it with such artificial TOR systems.

Results & Conclusions

Summary

Objectives of the work

- Topical Opinion Retrieval (TOR) is evaluated by classical IR evaluation measures, i.e. Mean Average Precision (*MAP*) or Precision at 10 (*P@10*).
- The effectiveness of the topical-only retrieval (effectiveness of the baseline) boosts the TOR performance.
- How can we assess the opinion-only classification accuracy (or precision, etc.)? How can we split the contribution of the opinion component from retrieval?

Methodological Framework

- We build artificial opinion-only classifiers from relevance and opinion data at different rates of opinion accuracy and precision.
- Then we study the effect on MAP of the TOR system with such classifiers
- We are able to assess the opinion-only component quality of a given TOR system by comparing it with such artificial TOR systems.

Results & Conclusions

Summary

Objectives of the work

- Topical Opinion Retrieval (TOR) is evaluated by classical IR evaluation measures, i.e. Mean Average Precision (*MAP*) or Precision at 10 (*P@10*).
- The effectiveness of the topical-only retrieval (effectiveness of the baseline) boosts the TOR performance.
- How can we assess the opinion-only classification accuracy (or precision, etc.)? How can we split the contribution of the opinion component from retrieval?

Methodological Framework

- We build artificial opinion-only classifiers from relevance and opinion data at different rates of opinion accuracy and precision.
- Then we study the effect on MAP of the TOR system with such classifiers
- We are able to assess the opinion-only component quality of a given TOR system by comparing it with such artificial TOR systems.

Results & Conclusions

The topical opinion retrieval (TOR)

TOR systems have two phases:

Topic Retrieval : Ranking documents by content-only;

Opinion Mining : Filtering or re-ranking these documents by opinion content.

Filtering or re-ranking relevant documents by opinions always hurts the initial performance of topical retrieval (with the actual TREC submitted runs). Actually MAP always increases with a perfect opinion classifier!

To assess the effectiveness of an opinion mining strategy should be sufficient to observe MAP of relevance and opinion ($MAP_{\mathcal{R},\mathcal{O}}$) with respect to MAP of the baseline.

Unfortunately different baselines provide different increment rates for the same technique of opinion mining.

To sum up

The aim of our work is to introduce a methodological evaluation framework to:

- provide a best achievable $MAP_{\mathcal{R},\mathcal{O}}$ for a given baseline;
- assess opinion mining effectiveness from the overall topical opinion retrieval performance;
- study best filtering strategies on top of topical retrieval.

Artificial opinion classifiers

Let A be a *complete* set of assessments (by topic-relevance and opinion-only) for the collection. A binary opinion classifier is a function that maps documents in $C_{\mathcal{O}}$, the category of opinionated documents, and $C_{\overline{\mathcal{O}}}$, the category of non-opinionated documents.

	$\mathcal{O}$	$\overline{\mathcal{O}}$
$C_{\mathcal{O}}$	$K_{\mathcal{O}} \cdot \lvert\mathcal{O}\rvert$	$(1 - K_{\overline{\mathcal{O}}}) \cdot \lvert\overline{\mathcal{O}}\rvert$
$C_{\overline{\mathcal{O}}}$	$(1 - K_{\mathcal{O}}) \cdot \lvert\mathcal{O}\rvert$	$K_{\overline{\mathcal{O}}} \cdot \lvert\overline{\mathcal{O}}\rvert$

We define a class of artificial binary classifiers of opinion, $C^{A}_{K_{\mathcal{O}},K_{\overline{\mathcal{O}}}}(\cdot)$, where

- $K_{\mathcal{O}}$ is the detection rate of true positive documents according to A;
- $K_{\overline{\mathcal{O}}}$ is the detection rate of true negative documents according to A;
- $(1 - K_{\mathcal{O}}) \cdot |\mathcal{O}|$ is the number of *type I* errors;
- $(1 - K_{\overline{\mathcal{O}}}) \cdot |\overline{\mathcal{O}}|$ is the number of *type II* errors;

How to use the framework

Given a topical opinion retrieval run and its $MAP_{\mathcal{R},\mathcal{O}} = r$ value, we obtain the set of all $K_{\mathcal{O}}$ and $K_{\overline{\mathcal{O}}}$ values, such that the artificial opinion classifiers $C^{A}_{K_{\mathcal{O}},K_{\overline{\mathcal{O}}}}(\cdot)$ achieve r.
We then compute accuracy, precision, recall and F-score of the opinion-only component as follows:

- $\text{Acc} = \frac{K_{\mathcal{O}} \cdot |\mathcal{O}| + K_{\overline{\mathcal{O}}} \cdot |\overline{\mathcal{O}}|}{|\mathcal{O}| + |\overline{\mathcal{O}}|}$
- $\text{Prec} = \frac{K_{\mathcal{O}} \cdot |\mathcal{O}|}{K_{\mathcal{O}} \cdot |\mathcal{O}| + (1 - K_{\overline{\mathcal{O}}}) \cdot |\overline{\mathcal{O}}|}$
- $\text{Rec} = K_{\mathcal{O}}$
- $\text{F-score} = 2 \cdot \frac{\text{Prec} \cdot \text{Rec}}{\text{Prec} + \text{Rec}} \ (\beta = 1)$

Any approach must improve the performance of the *random classifier* $C^{A}_{P(\mathcal{O}),1-P(\mathcal{O})}(\cdot)$, where $P(\mathcal{O}) = \frac{|\mathcal{O}|}{|\mathcal{C}|}$ is the a priori distribution of opinionated documents in the collection.

TREC Blog2008 collection

The Blog2008 consists of 3.2 millions of web pages containing blog posts, a test suite of 150 topics and a set of relevance/opinion assessment (QRELs).

Topics and QRELs are provided by the NIST.

The NIST also provided the best 5 runs, named baselines, produced by some participants. Each baseline is made by 150 runs, one for each topic.

Complete the data

Unfortunately the 150 topics are a sample of the topics treated by the collection and the largest part of documents are not assessed with respect to their content of opinion.

To fill the lack of information on the opinion expressed by documents we need to "complete" the data.

To complete the data we assume that each document is relevant for some topic t. *Qrels*$_t$ is completed assigning each non relevant document for t to the set of non relevant and opinionated documents with probability

$$P(\mathcal{O}_{\overline{\mathcal{R}}_t}) = \frac{|\mathcal{O}_{\mathcal{R}} - \mathcal{O}_{\mathcal{R}_t}|}{|\mathcal{R} - \mathcal{R}_t|}.$$

Analogously can be defined $P(\overline{\mathcal{O}}_{\overline{\mathcal{R}}_t})$ as:

$$P(\overline{\mathcal{O}}_{\overline{\mathcal{R}}_t}) = \frac{|\overline{\mathcal{O}}_{\mathcal{R}} - \overline{\mathcal{O}}_{\mathcal{R}_t}|}{|\mathcal{R} - \mathcal{R}_t|} = 1 - P(\mathcal{O}_{\overline{\mathcal{R}}_t}).$$

The Monte Carlo approach

We use Monte Carlo approach to generate randomly different opinion assessments for not relevant data in order to complete data.

We iterate previous step to generate randomly different values for precision, recall, F-score or accuracy and average them.

Much less than 20 cycles are enough to obtain stable results.

How to use the framework to predict opinion performance

Setting $K_{\mathcal{O}} = K_{\overline{\mathcal{O}}} = 1$ the framework works as an oracle and provides a best achievable $MAP_{\mathcal{R},\mathcal{O}}$ for each baseline.

	$MAP_{\mathcal{R}}$	$MAP_{\mathcal{R},\mathcal{O}}$	$MAP^{*}_{\mathcal{R},\mathcal{O}}$	$\Delta^{\%}$
BL1	0.3540	0.2639	0.4999	89%
BL2	0.3382	0.2657	0.4737	78%
BL3	0.4079	0.3201	0.5580	74%
BL4	0.4776	0.3543	0.6294	78%
BL5	0.4247	0.3147	0.5839	86%

Mean Average Precision of relevance $MAP_{\mathcal{R}}$, relevance and opinion $MAP_{\mathcal{R},\mathcal{O}}$, optimal relevance and opinion $MAP^{*}_{\mathcal{R},\mathcal{O}}$, variation $\Delta^{\%}$ between $MAP^{*}_{\mathcal{R},\mathcal{O}}$ and $MAP_{\mathcal{R},\mathcal{O}}$.

Mean percentage variations of $MAP_{\mathcal{R},\mathcal{O}}$ filtering the baselines through $C^{Qrels^*}_{K_\mathcal{O},K_{\overline{\mathcal{O}}}}(\cdot)$.

$K_{\overline{\mathcal{O}}}$ \ $K_\mathcal{O}$	1.0	0.9	0.8	0.7	0.6	0.5
1.0	81%	63%	45%	27%	10%	-9%
0.9	63%	46%	28%	11%	-7%	-24%
0.8	50%	33%	17%	0%	-17%	-33%
0.7	40%	24%	7%	-8%	-24%	-39%
0.6	32%	16%	0%	-15%	-30%	-44%
0.5	24%	9%	-6%	-20%	-35%	-48%

$K_\mathcal{O}$ contributes to improve $MAP_{\mathcal{R},\mathcal{O}}$ more than $K_{\overline{\mathcal{O}}}$. This is evident comparing the values of $MAP_{\mathcal{R},\mathcal{O}}$ reported by the column and the row corresponding to $K_\mathcal{O} = K_{\overline{\mathcal{O}}} = 0.7$.

Use the framework to compare the best three TREC approaches

The best three approaches to the TREC Blog Track 2008 achieve, on the five baselines, the following performance:

1. $MAP_{\mathcal{R},\mathcal{O}} = 0.3614$, percentage improvements of $+12\%$;
2. $MAP_{\mathcal{R},\mathcal{O}} = 0.3565$, percentage improvements of $+10\%$;
3. $MAP_{\mathcal{R},\mathcal{O}} = 0.3412$, percentage improvements of $+5\%$;

These evidently different improvements do not significantly differ in terms of opinion mining effectiveness.

Conclusion

- Our evaluation framework assesses the effectiveness of opinion mining techniques.
- This framework, makes it possible to provide a best achievable $MAP_{\mathcal{R},\mathcal{O}}$ for a given baseline.
- We determine the minimum values of accuracy, precision, recall and F-score that make it possible to improve a baseline. These values show that it is an hard task to improve a baseline by filtering its documents according to the opinion they express.
- We show how to compare different opinion mining techniques and to understand if they really improves on the state of the art.

Thanks!