Bootstrapping subjectivity detection

Jijkoun, V.; de Rijke, M.

DOI
10.1145/2009916.2010081

Publication date
2011

Document Version
Final published version

Published in
SIGIR'11

Bootstrapping Subjectivity Detection

Valentin Jijkoun*
ISLA, University of Amsterdam
jijkoun@uva.nl

Maarten de Rijke
ISLA, University of Amsterdam
derijke@uva.nl

ABSTRACT

We describe a method for automatically generating subjectivity clues for a specific topic and a set of (relevant) document, evaluating it on the task of classifying sentences w.r.t. subjectivity, with improvements over previous work.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Linguistic processing*

General Terms

Experimentation, Measurement

Keywords

Subjectivity, sentiment retrieval

1. INTRODUCTION

We address the task of detecting on-topic subjectivity in text. Specifically, we want to (1) tell whether a textual document expresses an attitude (positive or negative) towards a specific topic, and moreover, (2) to find where exactly in the document it is expressed (up to a phrase or at least a sentence). The first task is in the area of *sentiment retrieval*. The simplest approach here consist of two stages: first, we find texts that are on topic, then we filter out (or, rank low) those without attitude [3]. A more elaborate approach is based on the assumption that documents are mixtures of two generative components, one "topical" and one "subjective" [4]. In practice, however, these components are not independent: a word that is neutral w.r.t. one topic can be a good subjectivity clue for another (e.g., compare *hard copy* and *hard problem*). Noticing this, Na et al. [6] generate a topic-specific list of possible clues, based on top relevant documents, and use this list for subjectivity filtering (reranking). Furthermore, Jijkoun et al. [1] argue that such clues are specific not only to the topic, but to the exact target they refer to, e.g., when looking for opinions about a sportsman, *solid* is a good subjectivity clue in the phrase *solid performance* but not in *solid color*.

Jijkoun et al. [1] describe a method for learning such pairs (clue, target) for a given topic in an unsupervised manner, using syntactic dependencies between clues and targets. Kim et al. [2] also use syntactic relations to bootstrap a set of topic-specific clues and use them for detecting sentences containing on-topic sentiment. Note that the methods in [1, 2] also address the second task introduced above: finding the exact location of sentiment in documents.

We go beyond the subjectivity lexicon generation methods from [1, 2], with the goal of improving subjectivity spotting. We extend the method of [1] using bootstrapping (similarly to [2]). Unlike [1], we directly evaluate the performance on the task of detecting on-topic subjectivity at the sentence level, not on sentiment retrieval with entire documents. Unlike [2], our method does not use a seed set for a given topic: we only need a general purpose subjectivity lexicon, a topic and a set of (presumably) relevant documents.

2. METHOD

We start with a topic T (a textual description) and a set $R = \{d_1, \ldots, d_N\}$ of documents deemed relevant to T. The method uses a general-purpose list of subjectivity clues L (in our experiments, the well-known MPQA lexicon [9]). We will also use a large background corpus BG of documents of a similar genre, covering many topics beside T. We use the Stanford syntactic parser to extract dependency relations in all sentences in all documents. Our method outputs a set of triples $\{(c_i, r_i, t_i)\}$, where c_i is a subjective clue, t_i a subjectivity target and r_i a dependency relation between the two words. We interpret an occurrence of such a triple in a document as an indication of sentiment relevant to T, specifically directed at t_i.

Our method operationalizes a number of intuitions. First, we assume that a given topic can be associated with a number of related targets (e.g., opinions about a sportsman may cover such targets as *performance, reaction, serve,* etc.) and each target has a number of possible clues expressing attitude towards it (e.g., *solid performance*). We assume that clues and targets are typically syntactically related (e.g., the target *serve* can be a direct object of clue *to like*), and every clue has syntactic relations connecting it to possible targets (e.g., for *to like* only the direct object can be a target, but not the subject, a adverbial modifier, etc.).

Step 1: Initial clue scoring. For every possible clue $c \in L$ and every type of syntactic relation r that can originate from it in the background corpus, we compute a *clue score* $s_{clue}(c, r)$ as the entropy of words at the other endpoint of r in BG (normalized between 0 and 1 for all c and r). The clue score gives an initial estimate of how well (c, r) may work as a subjectivity clue. Here, we follow the intuition of [1]: targets are more diverse than other syntactic neighbours of clues.

Step 2: Target scoring. For every word $t \in R$ we determine its target score that tells us how likely t is an opinion target related to topic T. Our main intuition here is that targets are words that occur unusually often in subjective contexts in relevant documents. First, we compute $C_R(t) = \sum s_{clue}(c, r)$ for all occurrences of the syntactic relation r between words c and t in corpus R. Sim-

*Current affiliation: Textkernel BV, Amsterdam.

SIGIR'11, July 24–28, 2011, Beijing, China.
ACM 978-1-4503-0757-4/11/07.

Method				P	R	F_1
method of [1]				0.23	0.31	0.26
R	K	N	M			
$r+100$	4	10	50	0.42	0.13	0.20
$r+100$	4	20	50	0.45	0.17	0.25
$r+100$	4	30	50	0.35	0.26	0.28
$r+100$	4	40	50	0.32	0.29	0.30
$r+100$	4	50	50	0.20	0.30	0.24
$r+100$	4	60	50	0.19	0.32	0.24
$r+100$	4	70	50	0.14	0.35	0.20
$r+100$	4	40	30	0.32	0.21	0.25
$r+100$	4	40	40	0.32	0.23	0.27
$r+100$	4	40	50	0.32	0.29	0.30
$r+100$	4	40	60	0.30	0.29	0.29
$r+100$	4	40	70	0.29	0.30	0.29
$r+100$	4	40	50	0.32	0.29	0.30
100	4	40	50	0.27	0.22	0.24
r	4	40	50	0.21	0.17	0.19

ilarly, we compute $C_{BG}(t)$ for the background corpus BG. We view $C_R(\cdot)$ and $C_{BG}(\cdot)$ as (weighted) counts, and compute a parsimonious language model $p_R(\cdot)$ using a simple EM algorithm [5]. We also compute a language model $p_BG(\cdot)$ from counts $C_{BG}(\cdot)$ by simple normalization. Finally, we define the target score of a word t as the likelihood that the occurrence of t in R comes from $p_R(\cdot)$ rather than $p_BG(\cdot)$:

$$s_{tgt}(t) = \frac{\gamma \cdot p_{tgt}(t)}{\gamma \cdot p_{tgt}(t) + (1-\gamma) \cdot p_{BG}(t)}.$$

Step 3: Clue scoring. Mirroring Step 2, we now use target scores to compute better estimates for clue scores. Here, our intuition is that good subjectivity clues are those that occur unusually often near possible opinion targets for a given topic. The computation is similar to Step 2, with $s_{clue}(c,r)$ and $s_tgt(t)$ interchanged: we compute weighted counts, a parsimonious model and, finally, the updated $s_{clue}(c,r)$. Now, we iterate Step 2 and Step 3, each time updating $s_{tgt}(\cdot)$ and $s_{clue}(\cdot,\cdot)$, respectively, based on the values at the previous iteration. After K iterations we select N targets and M pairs (clue, relation) with the highest scores. We check which of the N targets co-occur with which of the M clues in R.

3. EXPERIMENTS AND RESULTS

We evaluate different versions of our method on the following sentence classification task: for a given topic and a list of documents relevant to the topic, we need to identify sentences that express opinions relevant the topic. We compute precision, recall and F-score for detection of relevant opinionated sentences.

In our experiments, we use the NTCIR-6 [7] and NTCIR-7 [8] Opinion Analysis datasets, containing judgements for 45 queries and 12,000 sentences.

In order to understand how the quality of relevant documents affects the performance of the method, we selected R to be (1) R_{100}: top 100 document retrieved from the NTCIR-6/7 English collection using Lucene, (2) R_r: only documents with at least one relevant (not necessarily opinionated) sentence as identified by NTCIR annotators, and (3) R_{r+100} the union of (1) and (3).

We also ran the method with different numbers of iterations (K), different number of selected targets (N) and selected clues (M). In all settings, the overall performance stabilizes at $K \leq 5$. Table 3 shows the evaluation results:

As one might expect, we see that reducing the number of selected targets (N) improves precision but harms recall. Changing the number of selected clues (M) has little effect on precision: since for detecting opinionatedness we combine clues with targets, noise in clues does not necessarily lead to drop in precision.

Overall, we notice that with in the best setting ($K = 4$, $N = 40$, $M = 50$) the method outperforms [1] (significantly, at p=0.05, using t-test). Performance of the method varies substantially per topic (F_1 between 0.13 and 0.48), but the optimal values for parameters are stable for high-performing topics (with $F_1 > 0.26$).

4. CONCLUSIONS

We have described a method for automatically generating subjectivity clues for a specific topic and a set of (relevant) document, evaluating it on the task of classification of sentences w.r.t. subjectivity, demonstrating improvements over previous work. We plan to incorporate more complex syntactic patterns in our clues (going beyond word-word relations) and study the effect of user feedback (which extracted targets are correct? which clues are indeed subjective?) with the view of implementing an interactive system.

Acknowledgements

This research was partially supported by the European Union's ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, CIP ICT-PSP under grant agreement nr 250430, the PROMISE Network of Excellence co-funded by the 7th Framework Programme of the European Commission, grant agreement no. 258191, the DuOMAn project carried out within the STEVIN programme which is funded by the Dutch and Flemish Governments under project nr STE-09-12, the Netherlands Organisation for Scientific Research (NWO) under project nrs 612.-061.814, 612.061.815, 640.004.802, 380-70-011, the Center for Creation, Content and Technology (CCCT), the Hyperlocal Service Platform project funded by the Service Innovation & ICT program, the WAHSP project funded by the CLARIN-nl program, and under COMMIT project Infiniti.

REFERENCES

[1] V. Jijkoun, M. de Rijke, and W. Weerkamp. Generating focused topic-specific sentiment lexicons. In *ACL '10*, 2010.

[2] Y. Kim, Y. Choi, and S.-H. Myaeng. Generating domain-specific clues using news corpus for sentiment classification. In *ICWSM '10*, 2010.

[3] Y. Lee, S.-H. Na, J. Kim, S.-H. Nam, H.-Y. Jung, and J.-H. Lee. KLE at TREC 2008 Blog Track: Blog Post and Feed Retrieval. In *Proceedings of TREC 2008*, 2008.

[4] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai. Topic sentiment mixture: modeling facets and opinions in weblogs. In *WWW '07*, pages 171–180, 2007.

[5] E. Meij, W. Weerkamp, K. Balog, and M. de Rijke. Parsimonious relevance models. In *SIGIR '08*, 2008.

[6] S.-H. Na, Y. Lee, S.-H. Nam, and J.-H. Lee. Improving opinion retrieval based on query-specific sentiment lexicon. In *ECIR '09*, 2009.

[7] Y. Seki, D. K. Evans, L.-W. Ku, H.-H. Chen, N. Kando, and C.-Y. Lin. Overview of opinion analysis pilot task at NTCIR-6. In *Proceedings of NTCIR-6*, 2007.

[8] Y. Seki, D. K. Evans, L.-W. Ku, L. Sun, H.-H. Chen, and N. Kando. Overview of multilingual opinion analysis task at NTCIR-7. In *Proceedings of NTCIR-7*, 2008.

[9] J. Wiebe, T. Wilson, and C. Cardie. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39:165–210, 2005.