The Dangerous Myth of the Star System

André Bittar, Luca Dini, Sigrid Maurel, Mathieu Ruhlmann
Holmes Semantic Solutions,
Grenoble, France
E-mail: { bittar, dini, maurel, ruhlmann }@ho2s.com

Abstract

This paper describes a large scale experiment aimed to detect the reliability of users when converting their written opinions on product into a numerical score (e.g. number of stars). The study shows that, due to a number of factors, such a judgment is highly unreliable and, confronted with a uniform gold standard, provides accuracy inferior to a state of art system for opinion detection (namely the Senti-Miner system based on HOLMES (Hybrid Operable platform for Language Management and Extensible Semantics)). In particular, we show that user judgements are strongly biased towards positivity: this might be due to some features of the user interface, but it is highly probable that users are generally more inclined towards positivity when lacking clear criteria to transform a text into a number. Whatever the reasons of this bias might be, it is evident that this behaviour, if confirmed on different data sets, could pose some doubt on evaluation experiments based on user generated assignments as well as on systems whose training comes from the same sources.

Keywords: Opinion mining, Sentiment Analysis, Evaluation

1. Introduction

In recent years we have observed two parallel trends in computational linguistics research and e-commerce development. On the research side, there has been an increasing interest in algorithms and approaches that are able to capture the polarity of opinions expressed by users on products, institutions and services. On the other hand, almost all big e-commerce and aggregator sites are by now providing users the possibility of writing comments and expressing their appreciation with a numeric score (usually represented as a number of stars). This generates the impression that the work carried out in the research community is made partially useless (at least for economic exploitation) by an evolution in web practices.
In this paper we describe an experiment on a large corpus which shows that the score judgments provided by users are often conflicting with the text contained in the opinion, and to such a point that a rule-based opinion mining system can be demonstrated to perform better than the users themselves in ranking their opinions.

2. The experiment

In summer 2013, the French giant of food distribution Carrefour launched an experiment for collecting customer opinions on a certain set of products. The idea is that the customer will benefit from a certain amount of "credits" for buying certain products under test, and, in case s/he writes an opinion on the http://monavislerendgratuit.com site, s/he will get even more credits. The initiative is certainly successful, as a few months after its launch, the web site was already offering hundreds of thousands of product reviews (precisely, at the moment where we performed web scraping (09/09/2013), 141,248 reviews for 245 products). The users were asked both to write some textual comments (in French) and to rate the products on a scale of 0-5 stars.
The experiment we carried out went along the following lines:

1. Download and convert all opinions: comment text and user assigned score;
2. Process all textual comments by a state-of-the-art opinion extraction system;
3. Manually compare automatically extracted opinions scores with user assigned ones in order to determine their respective reliability.

In this paper, we will describe i) the system which was used to perform automatic opinion extraction, ii) the manual annotation process, iii) the lessons learned from the evaluation of the results against the manually annotated corpus.

3. Senti-Miner

Senti-Miner is a derivation of Sybille 2.0, a system for opinion monitoring presented at DEFT07 (Maurel et al. 2007; see also Maurel & al. 2008; Maurel & al. 2009), where it ranked third among all competing systems and first among the industrial ones. Contrary to Sybille 2.0, which was based on XIP (Mokhtar & al. 2001), Senti-Miner is built on top of the HOLMES (Hybrid Operable platform for Language Management and Extensible Semantics) system. The basic assumption of HOLMES is that hybridization of different technologies is essential in order to achieve good performance in generic text mining and information extraction tasks. It is therefore based on a flexible processing model (very similar to the Stanford CoreNLP platform assumptions) where different annotators are disposed in a pipeline and where every annotator can benefit from the processing of all previous annotators. We adopted the general pattern whereby we plugged pairs of annotators with comparable functionalities into the pipeline, one based on statistical techniques (mostly supervised), and one based on manual configuration. The role of the linguist then becomes to correct the output of the statistical model on the basis of appropriate rules. For instance, HOLMES contains both a CRF-based named entity recognition module (Lefferty & al., 2001) and a correction module based on TokenRegexp

(Levy and Galen, 2006), a stochastic POS tagger and a linear pattern matching rule component, a MaltParser-based model for dependency parsing and a graph transformation-based component for detecting and correcting parsing errors.
Hybridization also goes the other way around, in the sense that the learning phase for all trainable HOLMES modules can fully benefit of the input of symbolic processors: for instance grammatical dependency relations can be used as features of CRF learning. This double hybrization is also very important in the case of application of boosting approaches, where the initial seed is represented by the output of manually created rules.

In the case of Senti-Miner, the basic HOLMES machinery has been enriched with a semantic analysis component, described in the next section.

3.1 Semantic Analysis as Graph Transformation

A long standing research trend in computational semantics (cf. (Sowa, 2008) for an exhaustive coverage of the literature) assumes that an optimal representation of natural language semantics can be achieved by using a graph representation. HOLMES' semantic layer (which includes the sentiment analysis module) is based on this assumption. Basically, predicates (in a first order logic sense) are represented by arcs connecting nodes, which correspond to entities detected in the text and enriched with specific semantic information. For instance, the sentence "Le **patient** est **réadressé** en **Service d'Orthopédie le 19.02.2012**." ("The patient was transferred to the Orthopedics Ward on the 19.10.2012.") is represented as shown in Figure 1.

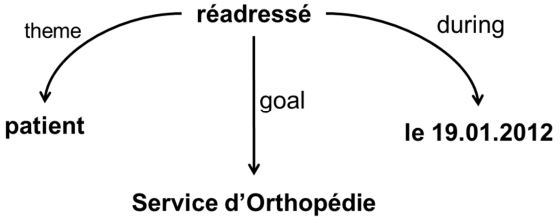

Figure 1: semantic representation output by HOLMES.

As the output of the "basic" HOLMES machinery is represented by a dependency graph, it is natural to conceive the process of semantic enrichment as a process of graph transformation, along the lines of Bonfante et al. (2010) and Ribeyre (2012). In our case this is further facilitated by the fact that the output conforms to the Stanford dependency paradigm (Manning & Marneffe 2008; Cer & al. 2010; Robin & al., 2013), i.e. with graph representations that are closer to a semantic representation than "standard" syntax-oriented dependency graphs.
The real challenge of conceiving semantic analysis as graph transformation lies in the fact that the rules governing the various transformation steps need to access a large amount of syntactic and lexical semantic information, whereas standard graph transformation platforms usually offer the possibility of handling limited alphabets. For this reason the natural orientation was towards attributed graphs as described in (Fisher & al., 1998), which find a natural implementation in the AGG environment (Taentzer, 2000). The software comes with a graphical user interface which allows the user to write rules for graph manipulation. Each rule describes a left input graph, a right output graph and a "host" graph where the transformation can be tested. As attractive as it may be, the AGG GUI is quite counter-productive for the intensive task of rule writing. This is probably because linguists are more oriented towards the writing of formal declarative rules than drawing arcs between objects. Moreover, there was the need, for reasons of efficiency and maintenance, to limit the formal power of AGG in such a way that in the application phase non-functional constraints of efficiency and computability were satisfied. For these reasons, we designed a declarative language for graph transformation. The language allows operations on graphs, such as the creation and the deletion of arcs and nodes, the declaration and assignment of attributes, etc. It also allows arbitrary calls to Java methods to test application preconditions and to assign functional values. A simple rule in our graph transformation language might look like the following:

```
[1][2 pos_pol=true pos=adj][3 pos=ADV
type=invert] {mod(1,2), mod(2,3)} =>
[1][2]{Negative_opinion(1,2)}
```

This rule states that if a positive polarity adjective modifies a noun, but is itself modified by a negative polarity adverb, then a negative opinion relation is established between the noun and the adjective (incidentally, the rule will also delete the adverb, as its contribution to semantics can be considered exhausted).
Rules of this kind constitute the core of the module which has been used for the experiment we described. The rules are divided into independent components (or layers), which apply in the order selected by the linguist. The module for French sentiment analysis currently contains 44 rules which access rich sentiment related lexical information.

4. The Annotation of the Corpus

Given the size of the corpus (141.248 reviews), a complete manual annotation according to the sentiment axis was not possible. Fortunately, however, this was not even necessary: the goal was not to produce a corpus for automatic learning (in which case the statement "the more the better" holds). The aim was to evaluate the level of consistency between the star ratings left by users with their corresponding textual evaluations. We proceeded then by discarding all the opinions where the verdict of Senti-Miner coincided with the user rating. As Senti-Miner produces a "positivity score"[1] from 0 to 1 for each opinion (1 being maximally positive, 0 being

[1] The score is the result of a weighting of all positive and negative expressions found in the text, together with their strength.

maximally negative), and as the human judgment is an integer between 1 and 5, it would have been normal to create equivalence classes such as 0-0.2=**1** 0.2-0.4=**2**, etc.(user-assigned scores are in bold) However, this five-dimensional classification was too fine-grained, and the vast majority of agreements emerged for the very positive opinions only. Therefore, we settled for a division based on three categories: definitely negatives (DN: 0-0.4/**1-2**), definitely positives (DP: 0.8-1/**4-5**) and undecided (U: 0.4-0.8/**3**). With these parameters we obtained a consensus [2] between Senti-Miner and user-assigned scores of 50% for DN, 76% for DP and 17% for U.

The manual annotation[3] of non-consensual opinions was performed with the human annotator having no clue about either Senti-Miner or user-assigned score. The instructions were simple: rank DP an opinion not containing any negative comment, as U any opinion containing a mix of positive and negative comments, and DN any text containing predominantly negative comments.
The following table summarizes the results in terms of precision, retrieval and accuracy of the Senti-Miner and user annotations when compared with the external human annotation, which is considered as the gold standard.

	Averaged precision	**Averaged retrieval**	**Accuracy[4]**
User	47%	37%	19%
Senti-Miner	45%	46%	59%

Table 1: Comparison between user scores and Senti-Miner scores.

5. Result Analysis

The figures for global precision and recall are actually only of partial interest. They just show that there are cases where a human judgment (the one provided by the user) is less reliable than the one provided by an NLP system. [5] What is more important is to try to interpret this finding by analyzing if user scores respond to a specific bias pattern. A comparison among the following two tables is quite enlightening in this regard:

	User prec.	User ret.	User acc.
DP	14%	95%	11%
DN	100%	6%	3%
U	27%	10%	4%

Table 2: User's precision compared to the gold standard

	SM prec.	SM ret.	SM acc.
DP	5%	4%	1%
DN	68%	70%	31%
U	61%	63%	27%

Table 3: Senti-Miner precision compared to the gold standard

As the results show there is a systematic bias of user score towards positivity. This is overtly evident for the DP class (with a retrieval of 95% for DP and 16% total for the remaining), and statistically significant for the other two classes (cf. the decrease of accuracy for U and DN). The bias is even more striking if one considers not the simple Boolean membership to a class, but the distance between them (a U assignment to a DN gold class should score better than a DP).

-

Irrespective of the conclusions, which will be drawn in the next section, we can explain this bias towards positivity in three ways (they do not necessarily apply all for the Carrefour test set):

1. The Graphical user interface itself has a bias towards positivity. The user just types her/his comments and leaves the default, which is always the highest score, active. [6]
2. The users, who, in the case of Carrefour, are rewarded for their opinions, feel that they stand to gain from a positive review. However, when describing the product in their comments, they stay closer to the real experience.
3. Finally, as there is no predetermined algorithm to convert opinions to scores, users always prefer to be on the positive side. [7]

By manually analyzing diverging data in the DP class we could identify that this guess is substantially confirmed on the ground of data. Indeed, there are basically three big classes of discordance between the gold standard and the user assignment:

1. Cases where the user simply assigned a positive judgment which is unrelated with his/her text.

[2] Here the consensus is simply computed as #_of_texts_assigned_by_SM_to_class_X/ #_of_texts_assigned_by user_to_class_X.
[3] The manual annotation concerned only the opinions where there was no consensus. As there were more than 30,000 of these, we took a sample and annotated only 3,000 non-consensual opinions. Sampling was done by considering the type of disagreement (e.g. user DN vs SM DP, user U vs SM DP, etc.), and maintaining their ration is the corpus to be annotated.
[4] Total correct / total
[5] This emerges in particular when considering the overall accuracy and the distance of rating.
[6] From a commercial point of view this is not surprising: vendor sites all have an interest in having positivity as a preference as this might increase theirs sales. The same holds for pure opinion sites such as www.productreview.com, www.ciao.com, as their business model is built on traffic towards merchant sites.
[7] This was evident in a previous experiment in the context of DEFT 2013 (Grouin et al. 2013). In that case the users who published a recipe on the web site http://www.marmiton.org/ were also asked to rate its difficulty. As mentioned in (Dini et al. 2013), given the absence of criteria to rank recipes, the bias towards positive rating ("very easy", in this case) was quite strong.

These cases cover reasons described in 1 and 2 and are exemplified by judgments such as:

> *je trouve que le nuggets est vraiment trop sec et la panure trop "croustillante". Les nuggets n`ont pas du tout plus à mon fils de 5 ans* [I find that the nuggets are really too dry and the bread-crumbs too crunchy. My 5 years son didn't like them at all.] (User score :5)
> *Ce produit avait une mauvaise consistance, écoeurante et désagréable car beaucoup trop molle et le goût était absent* [This product had a bad consistency, disgusting and unpleasant, as too soft and tasteless] (User score: 5).

2. Cases where a positive judgment is provided, but some feature is rated negatively: in these cases the users simply didn't take into account the negative part, thus disregarding completely the fact that s/he is offered a graded evaluation rather than just a positive/negative one. In this category of discrepancies we notice nevertheless that that there are features which are almost systematically disregarded if negative. A major one is price, as in:

   > *Un produit egficace* (sic !) *mais un peu cher, à achzter* (sic !) *dans le cadre d`une promo.* [An effective product, just a bit expensive in the context of a discount](user score 5)

3. Finally we find cases where the is a kind of implicit ellipsis, in the sense that the user assumes that all the positivity is encoded via the numeric score and only the "exceptional feature" (i.e. the negative one) needs to be reported, as for instance in this text about butter:

   > *L`emballage n`est pas pratique* [The packaging is not handy] (User score: 5).

   These cases are probably the worst ones from the point of view of the consequences that they could have, for instance, on the adoption of user generate opinions for training machine learning based systems: indeed, from the point of view of text only negative expressions are used to support a positive numeric judgment.

6. Further Works

Admittedly, the evidence reported in this paper is based on just one data source and some bias might be introduced by the specific settings of the web interface and by the fact that opinions are rewarded. It is therefore our intention to extend it to at least to two other review sites, possibly involving also other languages. In these new experiments we will also take into account available user information (place, age, activity on social media etc.) in order to verify if some correlation exists between structured score liability and other socio cultural features. It would also be extremely interesting to try to perform an automatic statistical analysis of the diverging cases to see if there are recurring patterns identifying the probability of a potential discordance between the numeric user-assigned score and the polarity of the text. We expect these patterns to involve both the content of the judgment (thus typically cultural in orientation, such as the feature "price" in the data-set under examination) and its syntactic expression (usage of certain adverbs).

7. Conclusions

In this paper we have shown on a substantial set of samples that user-assigned scores from Internet sources are often arbitrary and of inferior quality to the one assigned by a state-of-the-art opinion extraction system. This observation comes with two consequences, a positive one and a negative one.

The positive consequence is, of course, that research and experimentation in opinion monitoring and sentiment analysis does not risk being made obsolete by the expansion of user-generated scoring. On the contrary, the comparison between automatic and user-generated scores can provide even further insights into the analysis of a certain Internet population.

The negative consequence is that evaluation exercises on sentiment analysis which rely on automatically acquired scores (such as CAW 2.0-2009 and DEFT 2013, and unlike RepLab 2012 and 2013) should probably be revised and the conclusion drawn from them should be viewed with less certainty. A system scoring highly in such competitions is probably just very good at emulating user bias or errors.

8. References

Aït-Mokhtar S., Chanod J.-P. and Roux C. (2001). A multi-input dependency parser. In Proc. of IWPT'01.

Bonfante G., Guillaume B., Morey M., Perrier G., (2010). Réécriture de graphes de dépendances pour l'interface syntaxe-sémantique. In Proceedings of TALN'10, Montréal, Canada

Cer, D., de Marneffe, M-C., Jurafsky D., Manning, C. D. (2010). Parsing to Stanford Dependencies: Trade-offs between speed and accuracy. In 7th International Conference on Language Resources and Evaluation (LREC 2010).

Dini, L., Bittar, A., Ruhlmann, M. (2013). Approches hybrides pour l'analyse de recettes de cuisine. In Actes de DEFT. Cyril Grouin, Pierre Zweigenbaum, Patrick Paroubek, "DEFT2013 se met à table : présentation du défi et résultats" In Actes de DEFT.

Fischer I., Koch, M, Taentzer G., (1998), Visual Design of Distributed Sytems by Graph Transformation. Technical Report.

Levy R., Galen A. (2006). Tregex and Tsurgeon: tools for querying and manipulating tree data structures. In 5th International Conference on Language Resources and Evaluation (LREC 2006).

Lafferty, J., McCallum, A., Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of ICML

Manning C. D., Marneffe (de) M.-C., (2008), The Stanford typed dependencies representation. In Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation, pp.1-8

Maurel, S., Curtoni, P., Dini, L., (2007). Classification d'opinions par méthodes symbolique, statistique et hybride", actes du 3ème DÉfi Fouille de Texte, 2007, p. 111-117.

Maurel, S., Curtoni, P., Dini, L. (2008) "L'analyse des sentiments dans les forums", actes de l'atelier FOuille des Données d'OPinions, p. 9-21.

Maurel, S., Curtoni, P., Dini, L. (2009), « Sybille : anatomie d'un système automatique d'extraction de termes de sentiments », in Iva Novakova, Agnès Tutin (dir.), Le Lexique des émotions, Grenoble, ELLUG, pp. 275-296.

Ribeyre C., (2012). Mise en place d'un système de réécriture de graphes appliqué à l'interface syntaxe-sémantique. Université Paris Diderot 7, 73pp.

Robin C., Bittar A., Dini L. (2013). Stanford Dependencies for French: Annotation Guidelines and Dependency Transformation", to appear.

Sowa, J. F. Conceptual Graphs. (2008). In F. van Harmelen, V. Lifschitz and B. Porter, Handbook of Knowledge Representation, 2008 Elsevier, pp 213-237