Semantic Oriented Clustering of Documents

Alessio Leoncini[1], Fabio Sangiacomo[1], Sergio Decherchi[1], Paolo Gastaldo[1], and Rodolfo Zunino[1]

[1] SeaLab, DIBE, University of Genoa, via all'Opera Pia 11, 16135 Genoa, Italy
{alessio.leoncini, fabio.sangiacomo, sergio.decherchi, rodolfo.zunino}@unige.it

Abstract. Semantic web-based approaches and computational intelligence can be merged in order to get useful tools for several data mining issues. In this work a web-based tagging process followed by a validation step is carried to tag WordNet adjectives with positive, neutral or negative moods. This tagged WordNet is used to define a semantic metric for text documents clustering. Experimental results on movie reviews prove that the introduced semantically oriented metric is extremely fast and gives improved results with respect to the classical frequency based text mining metric from the accuracy point of view.

Keywords: Semantic Orientation; EuroWordNet; Clustering

1 Introduction

This work couples the *semantic orientation* task [1] with a clustering tool [2] in order to categorize and calibrate [3] documents. When referred to a word, semantic orientation (SO) indicates: "*the deviation of a word from the norm for its lexical field*" [1]; in this work the mood represents the semantic orientation; it can be either positive, neutral or negative. The paper shows that semantic orientation can be used for building a metric [3]. This task requires the evaluation of an average SO for every document, starting from a SO of each term inside the documents.

One of the tools used to support SO is EuroWordNet [4]: using adjectives, and tagging them with their mood allowed to obtain a mood-tagged version of EuroWordNet. The tagging process is human-based and was obtained with the generous help of many students of the University of Genoa [5]: the values obtained from the tagging process were evaluated and validated in order to avoid inconsistencies among EuroWordNet synsets with the goal of discriminating positive from negative or neutral feelings.

The paper is organized as follows: Section 2 describes the existing categorization framework on which this work is built and EuroWordNet semantic network; Section 3 describes how semantic orientation is embedded in the clustering engine, Section 4 discuss the experimental results obtained on a movie review database [6].

2 Text Clustering Engine

The reference text clustering engine is SLAIR [2]; SLAIR is a versatile framework whose software components are easily customizable. SLAIR essentially is clustering tool and is based on Kernel K-Means; additionally the usual pre-processing steps involved in text mining applications such as stemming, punctuation removal and stop words removal are available. SLAIR can also perform queries to EuroWordNet semantic network; these queries are used to obtain a semantic metric.

The text document representation for clustering used in SLAIR is the vector space model (VSM) [7]. Additionally SLAIR uses the classical frequency based metric coupled with a style based metric which allows improved clustering performances [3]. Concerning the text clustering process, it is based on the following phases:

1. Stemming is obtained by using the word morphing functions provided with original WordNet English package [8].
2. Stopwords removal is performed using a simple list of frequent words with no discriminative effects.
3. A vocabulary is built with all the stemmed words from all the analyzed documents.
4. If enabled, the frequency based representation is mapped to a semantic one [2] by using EuroWordNet. This feature is used in this work to obtain a SO score.
5. Finally Hierarchical Kernel K-means [9], with dynamic branch creation on the clusters tree is run.

2.1 The EuroWordNet semantic network

EuroWordNet [4] is a multilingual lexical database built by researches coming from several European countries.

The structure of EuroWordNet was inspired by the Princeton WordNet [10] project. Both EuroWordNet and WordNet contains many concepts belonging to all the lexical categories, or part-of-speech (POS): nouns, verbs, adjectives and adverbs.

The fundamental building block of these databases is the “synset”: a set of words which share the same meaning, e.g. {“car”, “auto”, “automobile”, “machine”, “motorcar”}. Synsets can be related to each other by semantic relations, and can belong to some hierarchies of concepts. According to the lexical category, synsets have different semantic relations.

Another very interesting feature of EuroWordNet is the Inter-Lingual Index (ILI), which has the purpose to provide an efficient mapping across languages. The EuroWordNet database for the English, is a mapping of the original WordNet 1.5 born in Princeton: hence the English database is the most fine-grained database and for this reason was also the base for the creation of the ILI.

The EuroWordNet semantic network allows to retrieve, for every term in a document, a synset with words linked in several ways to the queried term.

3 Semantic Orientation Tagging and Metric

Two are the main contributions of this work; the mood tagging of EuroWordNet and a mood guided metric for clustering. Next subsections elucidate on these aspects.

3.1 Tagging EuroWordNet

In order to achieve EuroWordNet tagging in positive/neutral/negative sensations, an offline preprocessing phase for the assignment of the labels for the semantic orientation was requested. Starting from the 112,641 synsets of EuroWordNet's English database, only those belonging to the adjective lexical category were chosen to perform tagging. The synsets containing words with the initial capital letter (e.g. those concerning nationalities) were considered as non taggable. After the selection of all the WordNet adjective synsets, 15,726 synsets were obtained; on-line [5] a human-based categorization in positive/neutral/negative sensations was made possible by the generous contribution of many students of the Engineering Faculty of the University of Genoa.

This approach leads to pros and cons: from a perspective, a web-based tagging allows a large number of users to tag adjectives, however the process is not supervised and inconsistencies among synonyms can appear. A first solution to address this problem is to repeat more times (i.e. three) the full tagging, thus obtaining redundant scores for every synset, and so minimizing, as much as possible, the subjectivity of the results. The manual check of randomly picked samples confirmed the goodness of the votes when corrected in such a way. Table 1 shows the answers obtained by the volunteers.

Table 1. Semantic Oriented percentual categorization based on students answers

Adjective synsets	Positive sensation	No sensation	Negative sensation
15,726	3045 (19.4%)	8966 (57.0%)	3715 (23.6%)

3.2 Validation of the tagging process

This first stage gave a preliminary tagged EuroWordNet. In order to further increase robustness of the tagging a validation stage was adopted.

The synsets filled with adjectives are mainly characterized by two semantic relations: "synonymy" (indication of same concept) and "antonymy" (indication of opposite concept). Hatzivassiloglou and McKeown [11] in their work on the prediction of SO, claim that "*if we know that two words are related to the same property but have different orientations, we can usually infer that they are antonyms*"; in the current work, the opposite problem is present: the EuroWordNet databases provide reliable antonym relations and on the basis of these relations one can validate the tagging process, hence eliminating possible inconsistencies among synsets.

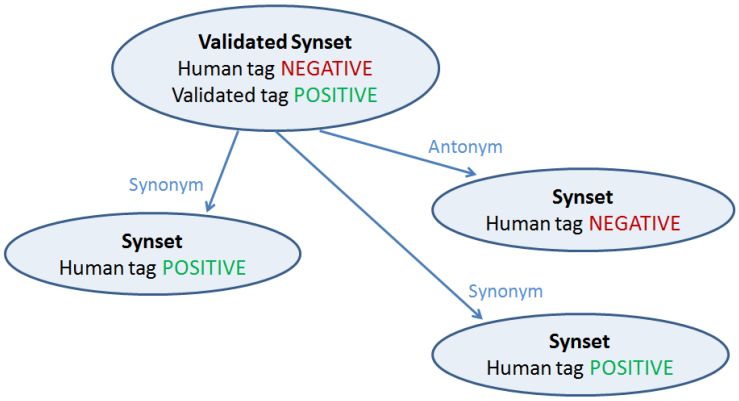

Fig. 1. Synsets validation example: the synset under test is inconsistent with respect to all its neighborhoods. Thus his tag is swapped accordingly from negative to positive.

The central idea on which validation is based is that redundancies allow to fix possible inconsistencies: operatively when a synset is under validation all its synonyms and antonyms are analyzed and the consistencies of the links among synsets are checked; if an antonym is found its vote is inverted, that is a positive SO becomes negative and vice-versa; reasonably a neutral vote remains neutral.

If the synset under test is not consistent with the majority of the relations to the other synsets, then its SO is set using majority voting induced by the other synsets. In this procedure one suppose that all the synsets surrounding the synset under analysis are correctly tagged: this assumption is approximately correct also for a roughly tagged network because a synset is usually linked to a large number of synsets and one can reasonable suppose that the high number of synsets allow a robust validation. Figure 1 graphically explicates the procedure; in that case a synset was humanly tagged in a inconsistent way with respect to the other synsets, thus its mood is changed accordingly.

The full procedure can be summarized in the following pseudo-code:

```
For each synset s with a vote v, create a vector v:
  Get from EuroWordNet the synonyms S of s;
    Append the vote v of each element of S to v;
  Get from EuroWordNet the antonyms A of s;
    Invert the vote v of each element of A obtaining v';
    Append the vote v' of each element of A to v;
End for.

For each synset s with an array of votes v:
  Counts v+,v-,v0: the number of positive, negative and
neutral votes;
  Find majority vote v* and assign it to s;
  If tie, the human vote wins;
End for.
```

After one validates the human-tagged adjectives one can further increase the number of tagged adjectives themselves. Indeed all the available synonyms and antonyms linked to the validated adjectives can be safely tagged. Results of the validation phase and further tagging process are showed in Table 2.

Table 2. Results from the WordNet processing and validation of students answers.

Adjective synsets	Positive sensation	No sensation	Negative sensation
15,842	4010 (25.3%)	7371 (46.5%)	4461 (28.2%)

From the Tab. 2 results, one sees that the number of concepts that express a positive sensation has become closer to the number of concepts expressing negative sensations with respect to Table 1; more importantly the number of neutral adjectives has been shrunk of 10.5%. This can be interpreted as a good result, as it is reasonable to suppose that each positive adjective roughly has its negative counterpart.

3.3 Semantically Oriented Metric

After the validation process each adjective synset exhibits a single vote, thus the votes can be used by the clustering tool for a metric.
In particular every document d_i is associated to a quantity m_i that is the mean of the adjectives mood present in the document, where a positive adjective is encoded with +1, negative with -1, and neutral with 0, thus m_i is in the range [-1,+1]; hence m_i represents the mood of the document d_i.
This mood m_i can be parameterized by adopting a strategy reminiscent of an *activation function* $f(m_i)$; $f(m_i)$ is a function that using a threshold τ can mitigate or emphasize the positive and negative mood. The function $f(m_i)$ is parameterized by the threshold τ and is defined as:

$$f(m_i) = \begin{cases} +1, \text{ if } m_i > \tau \\ m_i, \text{if } m_i \in [-\tau, +\tau] \\ -1, \text{if } m_i < -\tau \end{cases}$$

In the limit case for which $\tau = 0$, one obtains hard thresholding that is $f(m_i)$ is a binary function and the only available values are +1 and -1; this limit case enforces a very crisp solution in which positive and negative tagged documents are highly unrelated. The opposite case is the value $\tau = 1$, for which due to the range [-1,+1] of m_i then $f(m_i) = m_i$ thus the documents mood varies smoothly. Once the $f(m_i)$ scalars are computed these induce a simple Euclidean metric: given d_i and d_j the distance $D(d_i, d_j)$ is $(f(m_i) - f(m_j))^2$.
This parameterization allows to define different metrics that induce different clustering results where the concentration of the documents varies: in the limit case τ

= 0, at least in theory, only two clusters are possible the positive and the negative one; in the other limit case $\tau = 1$ clustering tool is free to get any result. To some extent the role of τ is that of a regularizer which defines the complexity of the solution, i.e. the number of clusters; the more the metric is concentrated (small τ) the less the number of clusters.

4 Experimental Results

The aim of this section is to compare the developed metric with the default SLAIR metric [3] within a suitable text mining domain. Being the object of this study the mood, or the affective side in texts, a suitable domain is offered by the movie review database [6]; the database used is the *polarity dataset v1.0* freely available at http://www.cs.cornell.edu/people/pabo/movie-review-data/. It collects 1400 movie reviews that are pre-classified in positive and negative reviews; the purpose of the experiments is to understand if a semantically oriented metric can be used instead of the usual text frequency metric in a context where the goal is classifying documents in two specific categories, positive and negative ones. Another aspect that is assessed is the influence of the threshold τ on the clustering results.

Table 3 compares the default SLAIR metric with the SO oriented metric with varying threshold values. The comparison is performed in terms of number of clusters, accuracy of classification and time needed to perform clustering.

The first aspect that emerges is that the new metric is extremely fast; indeed computing a distance value between 2 documents only requires a difference and a square, opposed to the classical frequency based distance that is much more expensive. In particular one obtains that the semantic metric is at least one order of magnitude faster than the usual metric used in text mining.

From the accuracy point of view the newly defined metric is very effective too; indeed with only 64 clusters accuracy is higher than that obtainable by the usual metric and 149 clusters.

Experiments also make clear the role of the threshold τ; when $\tau=0$ only 3 clusters are obtained; the first has size 648, the second 750, and the last 2. This means that the clustering engine has split the data almost exactly in two clusters thus giving an highly crisp result.

For increasing values of τ both the number of clusters and accuracy grow accordingly, meanwhile execution time is not affected by the augmented number of clusters: this happens because the total clustering time is dominated by distance computation and not by Kernel K-Means iterations. Finally when $\tau=1$ the accuracy is 70.36% that is higher than the SLAIR baseline metric that achieves 67.71%.

Table 3. Comparison between default SLAIR metric and semantic oriented metric parameterized by τ

Metric	τ	Time	Accuracy	Number of Clusters
Default	**-**	**9531 ms**	**67.71%**	**149**
SO	**0**	**656 ms**	**64.86%**	**3**
SO	0.01	562 ms	65.50%	9
SO	**0.1**	**562 ms**	**69.00%**	**64**
SO	0.2	578 ms	69.57%	101
SO	0.5	672 ms	69.43%	136
SO	0.75	610 ms	69.93%	134
SO	**1**	**562 ms**	**70.36%**	**139**

5 Conclusions

The proposed work shows how an affective technique called semantic orientation can be embedded into a clustering tool for text mining. A preparation phase involved many students from the University of Genoa to build, in a social-based way, a database containing couples adjective-sensation. To validate the obtained results, a method was proposed to improve consistencies of the first tagging stage that robustly corrected and validated the first human-based process.

Semantic Orientation was used to define a metric for clustering. Obtained results showed that the metric is extremely fast and gives improved accuracy with respect to the usual text frequency metric.

Future works will study other variants of the metric and other semantic orientations.

References

1. Lehrer, A.: Semantic fields and lexical structure. North Holland, Amsterdam and New York (1974)
2. Sangiacomo, F., Leoncini, A., Decherchi, S., Gastaldo, P., Zunino, R.: SeaLab Advanced Information Retrieval. Proc. of IEEE International Conference on Semantic Computing, 444--445 (2010)
3. Decherchi, S., Gastaldo, P., Zunino, R.: K-Means clustering for Content Based Document Management in Intelligence. Advances In Artificial Intelligence for Privacy Protection and Security. Solanas, A., Bellesté, A.M. (eds.) World Scientific Publishing (2009)
4. Vossen, P.: EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Kluwer Academic Publishers, Dordrecht (1998)
5. SeaLab Natural Language Processing Test, http://effimero-esng.dibe.unige.it
6. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment Classification using Machine Learning Techniques, Conference on Empirical Methods in Natural Language Processing (2002)
7. Salton, G., Wong, A., Yang, L.S.: A vector space model for information retrieval. Journal Amer. Soc. Inform. Sci. 18, 613--620 (1975)

8. WordNet morphology functions, http://wordnet.princeton.edu/wordnet/documentation/
9. Zhang, R, Rudnicky, A.: A large scale clustering scheme for kernel k-means, In: ICPR'02, 289--292 (2002)
10. Miller, G.A.: WordNet: A Lexical Database for English. Communications of the ACM 38(11) 39--41 (1995)
11. Hatsivassiloglou, V., McKeown, K.R.: Predicting the semantic orientation of adjectives. Proc. of 35th Annual Meeting of the ACL, 174--181 (1997)