Semantic Propagation on Contextonyms using SentiWordNet

Ovidiu Şerban*,** Alexandre Pauchet* Alexandrina Rogozan* Jean-Pierre Pécuchet*

*LITIS, INSA de Rouen,
76801 Saint-Étienne-du-Rouvray, France
email: {firstname.surname}@insa-rouen.fr

**Computer Science Department
"Babeş-Bolyai" University
Cluj-Napoca, Romania

Abstract:
Sentiment analysis and affect detection algorithms are generally based onto annotated data, structured into dictionaries, ontologies or word nets. The focus, so far, has been concentrated on manual annotation of the data, and then, in some situations, a semantic valence propagation is applied. The problem with this approach is that while it is able to build new affective labels through the propagation process, the precision of the decision decreases. Our approach disambiguates through the data, by offering a strong context using a contextonym model, for the usage of a certain term with a valence.

Keywords: Valence Disambiguation, Semantic Valence Propagation, Contextonyms

1 Introduction

In the field of sentiment analysis and emotion detection based on text data, two main directions for research exist : one concentrating on building better annotations of linguistic resources, such as dictionaries or ontologies, and the other on building better classifiers for valence, sentiment or emotion detection [4]. Very often, building a classifier, relies on having good linguistic resources. Improving these dictionaries, by increasing their size and annotation accuracy, is therefore considered mandatory in this field.

Unfortunately, even if recent approaches have increased the size of the dictionaries, the ambiguity of the decision increased as well [19]. Our goal is to improve these dictionaries by preserving, as much as possible, the annotation accuracy. This objective is performed by taking into account the context of the word, and a new linguistic approach to model this relation, called the contextonym model.

WordNet (WN) or variations over it, remain one of the most used linguistic resources, so far [4]. WordNet [10] is a lexical resource, build at Princeton University, which is mainly used in most of the Natural Language Processing (NLP) applications. The concepts in WN are grouped into synonym sets (also called synsets), which are sets of words semantic linked. Each synset may contain its frequency in the dictionary, and a gloss, which is basically a short sentence describing the sense of the synset. Among the basic synonymic relations, the WN contains also some special relations called : hyperonymy, hyponymy or ISA ("is a"). All these links describe generalisation, specialisation or equivalence relationships between some concepts. All these special links are introduced in WordNet 3.0 for a part of the synsets.

As a synset database example, we mention WordNet Affect [26], an extension of the WordNet [17] data set. WordNet Affect is basically a 6 class emotional annotation (i.e. Ekman's basic annotation scheme [8] : Anger, Disgust, Fear, Happiness, Sadness and Surprise) made on a synset level. It contains nouns, adjectives, adverbs and some verbs for the English WordNet 2.0 version.

ConceptNet [14] is another well-known ontology used widely for semantic disambiguation in classification tasks. This database contains assertions of common-sense knowledge encompassing the spatial, physical, social, temporal, and psychological aspects of everyday life. ConceptNet was generated automatically from the Open Mind Common Sense Project [22].

Another database, used especially for opinion and valence classification, is SentiWordNet (SWN) [2]. Valence is represented by the degree of positivity, negativity or neutrality of a certain word or sentence, while opinion represents the general valence over multiple sentences. SWN is the result of a semantic propagation algorithm over all WordNet synsets according to their valence. This resource is presented in more details in the following sections.

Starting from WordNet Affect, Valitutti et al. [28] proposed a simple word presence method

in order to detect emotions, where the emotion of a sentence is given by the dominant word emotions. Ma et al. [15] designed an emotion extractor from chat logs, based on the same simple word presence. SemEval 2007 (task 14) [25] presented a corpus and some methods to evaluate it, most of them based on Latent Semantic Analyser (LSA) [7] and WordNet Affect presence. This corpus is the one used in our experiments, because it offers a consistent annotation with our approach.

Methods more related to signal processing were proposed by Alm et al. [1], Danisman et al. [5], or D'Mello et al. [6], based on feature extraction, selection and different classifiers. Alm et al. [1] used a corpus of child stories and a Winnow Linear method to classify the data into 7 categories. Using the ISEAR [30] dataset, a very popular collection of psychological data recorded around 1990, Danisman et al. [5] used different classifiers like Vector Space Model (VSM), Support Vector Machine (SVM) or Naive-Bayes (NB) method to distinguish between 5 categories of emotions.

In the following sections we present SentiWordNet (Section 2), with the current problems concerning this WordNet (Section 2.2). In the next section, we introduce the contextonym model, by presenting its advantages. In Section 4 our approach, based on a contextonym model to disambiguate over conflicts in SentiWordNet, is presented, followed by a usage example (Section 5). Finally, we conclude this article by presenting the benefits of our model and highlighting some ideas for future work.

2 SentiWordNet

Among other WordNet extensions, like MultiWordNet [18], Balkanet [23], EuroWordNet [29] or WordNetAffect [26], SentiWordNet (SWN) [9] has been built as a lexical resource to help the valence prediction of a sentence. Its main field of application is opinion mining or sentiment classification. This resource contains annotation for mainly all the WordNet 3.0 synsets, having for each link a degree of positivity, negativity or objectivity annotated. Each of these valences are defined on a scale from 0.00 to 1.00, with the sum of all three of them being also 1.00.

Figure 1 presents the word **"good"** annotated according to SentiWordNet. For the selected synset, good has a definite positive value, of 1.0. The authors of SWN propose a triangle based visualisation, where each corner represents a different side of the valence : Positive (P), Negative (N) and Objective (O).

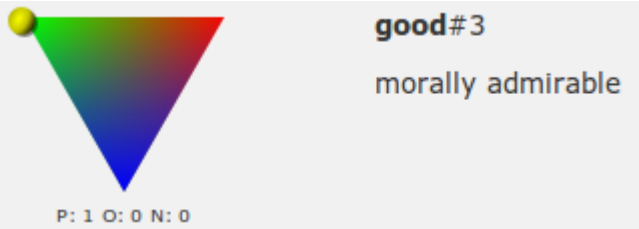

FIGURE 1 – A SentiWordNet [9] example, containing also the visualisation model. P states the positive degree, N the negative and O the objective one

2.1 Semantic valence propagation

The method of semantic valence propagation refers to diffusion of a valence through a structured corpus. The spreading is done by respecting the structure links between the words. The structures tested so far are different versions of WordNet, using the synsets. In a very frequent scenario, the process starts with a manually annotated set of words (also called "seeds"), and on every iteration the valence of these seeds is spread on the network. Examples of such approach are given by Rao et al. [20], Esuli et al. [9] and Godbole et al. [11]. Christopher Potts gives a more formalised definition of these algorithms in the Sentiment Analysis Tutorial [19], on the Semantic Valence Propagation section.

A more complex approach involves weighted propagation of valences [3], by not only spreading the valences in the neighbourhood of the seed nodes, but also computing a ranking measure attached to a node. This ranking measure is further used as a weight, taking into account the neighbourhood density among with node frequency.

Among others, SentiWordNet is the largest dictionary generated using valence propagation. However, SentiWordNet could not be manually reviewed for a better accuracy, mainly because of its size and the fact that WordNet structure does not disambiguate well between multiple usages. Given so, the valence conflicts remain unfortunately active in the database.

2.2 SentiWordNet and context

Christopher Potts [19] conducted an inconsistency level study between several opinion mining resources, and the results are presented in Table 1. This disagreement level between SentiWordNet and other corpora is due to the construction of SentiWordNet (based on automatic semantic propagation) and its size. Compared to the largest manually annotated dictionary, Harvard General Inquirer [24], SentiWordNet has almost 10 times more annotated words, which would give a better coverage over WordNet.

Dictionary	Dis. [a]	Annot. [b]	Cnt. [c d]
MPQA [31]	27%	+/-	8,221
Op. Lexicon [13]	25%	+/-	6,789
Gen. Inq. [24]	23%	+/-	11,788
LIWC [27]	25%	Categ. [e]	4,500

a. Disagreement level according to Potts [19]
b. Annotation style available in the corpus
c. Word countm in the corpus
d. SentiWordNet word count is 117,659
e. Words are grouped in several psychometric categories

TABLE 1 – Disagreement level between SentiWordNet and several other corpora

Table 1 presents an average of 25 % disagreement between SentiWordNet and MPQA [31], Opinion Lexicon [13], Harvard General Inquirer [24] or LIWC [27].

MPQA (Multi-Perspective Question Answering) Subjectivity Lexicon is a resource maintained by Theresa Wilson, Janyce Wiebe, and Paul Hoffmann [31] and contains annotations based on the subjectivity level, part of speech and polarity. Polarity corresponds to a discrete valence annotation, having a label for positive or negative.

Opinion Lexicon is maintained by Bing Liu [13] and contains discrete manual annotations for positive and negative words.

Harvard General Inquirer [24] is a lexical resource which is concentrated in attaching syntactic, semantic and pragmatic information to part-of-speech tagged words. It contains positive, negative and hostile labels for most of its containing words.

Linguistic Inquiry and Word Counts (LIWC) [27] is a proprietary database, containing categorised words to their psycho-semantic state, which can be translated into negative or positive labels.

Another aspect of the problem is represented by the conflictual valences which describe the same word. There are two distinct situations :

1. the word has different valences among different synsets,
2. the word has conflictual valences in the same synset.

The first issue is partially solved. Considering that each synset corresponds to a certain usage (context) of that word, then the valence from SentiWordNet corresponds to that context. In practice, finding the proper context, only by using WordNet synsets, is quite challenging.

On the contrary, the second type of conflict is an artefact of semantic propagation algorithm and it is not resolved in SentiWordNet. A term, as part of the same synset, should not have opposing valences because it would lead to ambiguous decisions. In practice, this problem is similar to the first case, because a term with conflictual valences would have two different contexts, even if it is part of the same synset.

In the following example, we choose the '*heart*' synsets to present the two situations :

1. spirit#8 **heart#6** : *an inclination or tendency of a certain kind ; "he had a change of heart"*, **+0.5**
2. **heart#1** bosom#5 : *the locus of feelings and intuitions ; "in your heart you know it is true"*, **-0.125**
3. spunk#2 nerve#2 mettle#1 **heart#3** : *the courage to carry on ; "you haven't got the heart for baseball"*, **+0.25 -0.25**

Example 1 and 2 show an inter-synset ambiguity, because the valence of **heart** in example 1 is definitely positive, while in example 2 is negative. In the third example, the ambiguity is showed for the same synset.

Both of the conflictual cases state well that WordNet synsets are not the most adapted solution to describe context.

3 Contextonyms

Contextonyms were introduced by Ji et al. [12] in order to model a more flexible lexical structure, to be used in machine translation. Similar to synonyms, the contextonym model links

words, but instead of having an equivalence relation between them, context is modelled by observing the word co-occurrences in a certain window [1]. A graph-based structure is generated, having the words as nodes and the co-occurring frequencies as edges. In order to model a strong relation between the words, a clique exploration algorithm is usually applied. In the end, the cliques correspond to a strong context, which give a structure called "contextonym" [12].

In graph theory, a clique is represented by a complete sub-graph. In other words, it is a structure where every node is connected to all the other nodes part of this structure. Maximal cliques, represent the largest complete subgraph that could be generated by the selected set of nodes. In information retrieval, cliques represent a strong link between the words that are part of it, and this structure could be exploited as a context [16, 12].

In Figure 2 a full example of the contextonym neighbourhood is given, being centred around the word '*heart*'. The contextonym model has been trained using a subtitle corpus and annotated using the valences from SentiWordNet. The technical details of this approach are given in the following sections.

In order to reduce the noise, several filtering techniques are proposed by Ji et al. [12], which are also considered parameters of the resulting structure :
– a global filter, which eliminates all the nodes that occur very rarely in the corpus
– a local filter, which is applied to every node and remove the neighbours with a low occurrence
– a children filter, which is similar to the local filter but it is applied to the neighbours of every nodes

In our approach, we applied only a global filtering technique, by removing very low occurrences from the graph. The other two filters are integrated in the clique ranking measure, described in more details in the next sections.

3.1 Our solution : Semantic valence propagation and contextonyms

Once the contextonym model is built, the annotated labels could be spread. In our approach, the labels are represented by the valences extracted from SentiWordNet. We consider that each contextonym could not have conflictual valences (multiple values for the same word or opposite valences inside the same contextonym). In the case of a conflict, these are solved by choosing a single value for each conflictual word. In Figure 2, the labels are coloured according to their valence : blue for positive, red for negative, purple for mixed-value and light-grey for neutral.

4 Experiments

Our technique requires a multi-step process, each step assuring the output for the next phase. The first step, also called preprocessing, consists in the filtering and cleaning the text information. After this step a clique exploration is applied using the DDMCE algorithm [21]. This algorithm can be used for clique exploration on large and dynamic data, like semantic approaches and social networks. The third step consists in aligning a certain phrase (consisting in a bag of words) to a set a cliques. In order to keep the alignment process consistent, a ranking measure is proposed.

Building the linguistic model for the contextonyms extraction is the most difficult and important step. In order to keep a link to an actual spoken language, we decided to use a subtitle corpus, collected from multiple sources [2]. Finally, a total of 53,384 unique movie files were kept. Also, in order to keep our linguistic model clean, we have kept the best subtitle file for each movie title, using an author and download count filtering.

Preprocessing Step.
A preliminary step, specific to sentence tokenizing on subtitle files was applied, filtering all the time synchronisation data. Even if the SubRip [3] format is clean and simple, a template validation had to be done, in order to verify the integrity of the data extracted. Recently, most of the subtitles tend to have advertisements included at the beginning or at the end of the file, a brute filtering approach was applied in order to remove them.

1. The size of the window has been fixed to 5, after applying the filtering process

2. The corpus represents a part of the subtitle database from `http://www.opensubtitles.org/` and `http://www.podnapisi.net/`

3. SubRip (.srt) is a very basic text format, used to encode subtitle files

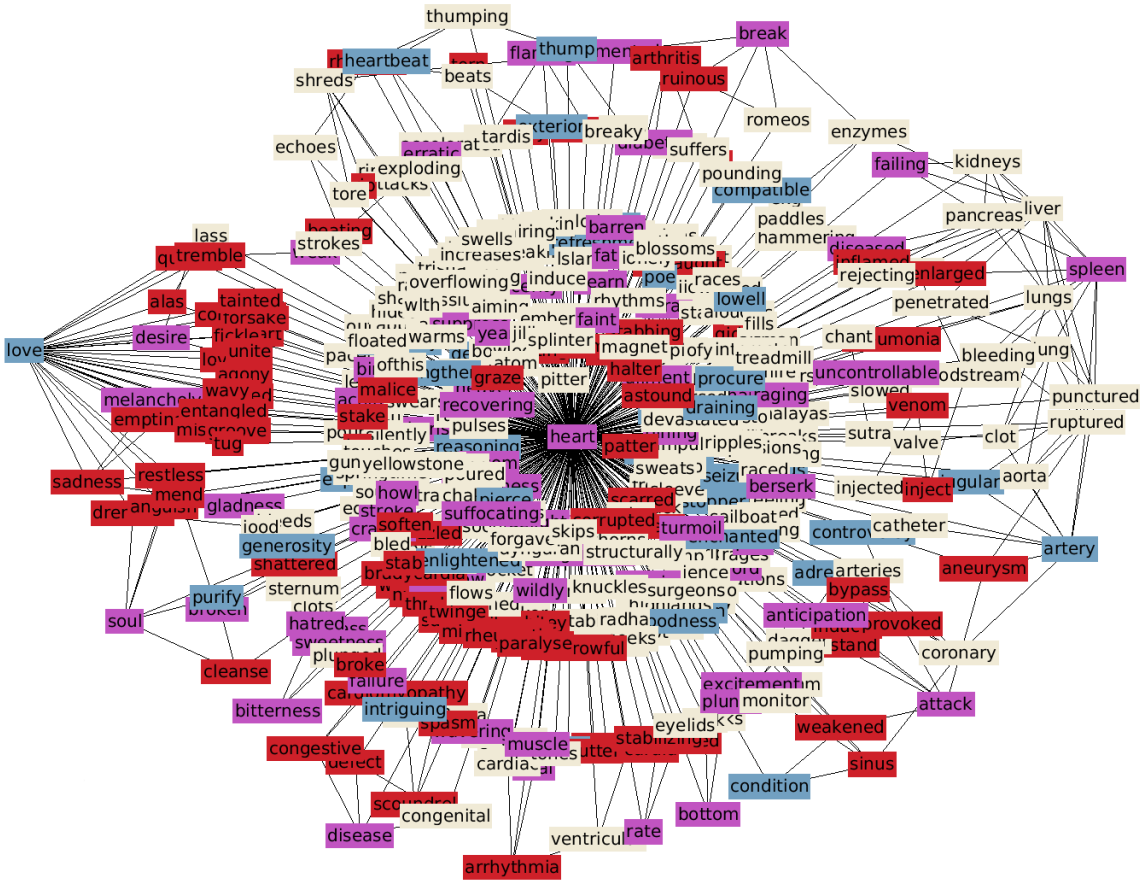

FIGURE 2 – A fully annotated contextonym graph, representing the whole neighbourhood of the word '*heart*'

During the preprocessing step, we applied on each headline a collection of filters, in order to remove any useless information, such as special characters and punctuation, camel-case separators and stop words. We considered as stop words, all prepositions, articles and other short words that do not carry any semantic value. The stop word collection used in our experiment is available at : `http://www.textfixer.com/resources/common-english-words.txt`. From the space reduction perspective, we kept only the words that are considered to carry a strong semantic and emotional value (e.g. nouns, verbs, adverbs and adjectives), as WordNet Affect is suggesting [26].

The method of stop word filtering offers a good balance between speed and accuracy of the results, compared to other methods like Part of Speech Tagging (POS), which provides comparable results, but tends to be much slower.

Clique Extraction.
After the preprocessing step, 86,276 words (frequency > 0.01 %) and 3,948,359 co-occurrences are maintained after filtering (frequency > 0.01 %). This was done to preserve most of the words collected, by limiting also the size of the problem. On this data, the DDMCE [21] algorithm was applied, obtaining a total of 702,546 cliques, from which 354,109 (50 %) are considered conflictual.

5 Usage example

The chosen corpus for our experiment is the one from SemEval 2007, task 14 [25], proposed at the conference with the same name. We believe both of the corpora used in our experiments : subtitles corpus and SemEval, are linked with oral data. The data set contains headlines (newspaper titles) from major websites, such as New York Times, CNN, BBC or the search engine Google News. The corpus was manually annotated by 6 different persons. They were instructed to annotate the headlines with emotions according to the presence of affective words or group of words with emotional content.

A valence annotation was carried out. Valence, as used also in psychology, means the intrinsic attractiveness (positive valence) or aversiveness (negative valence) of an event, object, or situation. In SemEval task, the valence is used to describe the intensity of the positive or negative emotion. The valence label ranged from -100 to 100.

From this corpus, we choose the item 24 : "*Hurricane Paul Weakens To Tropical Storm*", as an example for our approach. The complete approach can be observed in Figure 3. After the first parsing step, it could be observed that the word "tropical" is ambiguous. Given so, we selected the contextonym for tropical, exemplified in the Figure 4, and we try a best alignment with the existing context.

In our example, the alignment is made between tropical storm and cyclone (hurricane). This is a full alignment, since the word hurricane has the word cyclone as a direct hypernym (a generalisation). Doing this alignment, the disambiguation can be done, and the value decided for tropical would be negative.

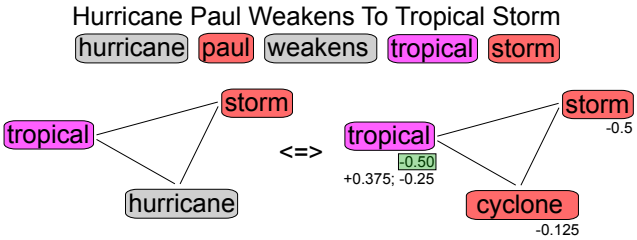

FIGURE 3 – Clique alignment for a SemEval 2007, task 14, item 24

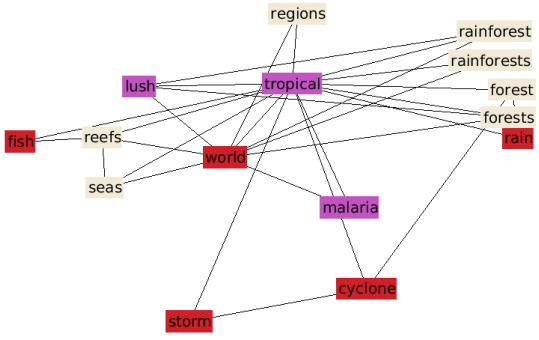

FIGURE 4 – The contextonym of the word tropical

5.1 Ranking measure

For a given sentence *ph*, corresponding to a set of words, equation 1 describe a ranking measure, used to discriminate partial or full alignments.

$$\forall q \in Q, R(q, ph) = \frac{f(q \cap ph) - f(q \setminus ph)}{f(ph)} \quad (1)$$

Where $f(X)$ represents the combined frequency of the set X. Q is the set of all possible cliques, and q is a clique from this set.

This measure is built to be used for partial alignments, but penalising the ones that are much larger than the actual sentence.

In the context of the previous example (Figure 3), the value of $R(\bullet, \bullet) = 0.3591$

6 Conclusion

In our approach, the context is a key part of the solution of the disambiguation problem. We model the context by using graph-based structures and we extract the strong-context by modelling it into contextonyms. By using these approaches, the disambiguation process is easy and more natural than in any previous work.

For the perspectives, a full validation has to be conducted. So far, only a small manual validation has been done. One of the major obstacles in the way of our validation process is the lack of free annotated resource that could be used.

A second perspective would be the integration of multiple modalities in our approach, like gestures, vocal features and other semantic approaches.

Acknowledgements

The work of this paper would not be possible without the help of the `www.opensubtitles.org` and `www.podnapisi.net` administrators. We thank them for their kindness of sharing a part of the database with us.

Références

[1] C.O. Alm, D. Roth, and R. Sproat. Emotions from text : machine learning for text-based emotion prediction. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 579–586. Association for Computational Linguistics, 2005.

[2] S. Baccianella, A. Esuli, and F. Sebastiani. Sentiwordnet 3.0 : An enhanced lexical resource for sentiment analysis and opinion mining. In *Seventh conference on International Language Resources and Evaluation, Malta. Retrieved May*, volume 25, page 2010, 2010.

[3] S. Blair-Goldensohn, K. Hannan, R. McDonald, T. Neylon, G.A. Reis, and J. Reynar. Building a sentiment summarizer for local service reviews. In *WWW Workshop on NLP in the Information Explosion Era*, 2008.

[4] R.A. Calvo and S. D'Mello. Affect detection : An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, pages 18–37, 2010.

[5] T. Danisman and A. Alpkocak. Feeler : Emotion classification of text using vector space model. In *AISB 2008 Convention Communication, Interaction and Social Intelligence*, volume 1, page 53, 2008.

[6] S.K. D'Mello, S.D. Craig, J. Sullins, and A.C. Graesser. Predicting affective states expressed through an emote-aloud procedure from AutoTutor's mixed-initiative dialogue. *International Journal of Artificial Intelligence in Education*, 16(1) :3–28, 2006.

[7] S.T. Dumais. Latent semantic analysis. *Annual Review of Information Science and Technology*, 38(1) :188–230, 2004.

[8] P. Ekman. Basic emotions. 1999.

[9] A. Esuli and F. Sebastiani. Sentiwordnet : A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, volume 6, pages 417–422, 2006.

[10] C. Fellbaum et al. Wordnet and wordnets. *Encyclopedia of Language and Linguistics*, pages 665–670, 2005.

[11] N. Godbole, M. Srinivasaiah, and S. Skiena. Large-scale sentiment analysis for news and blogs. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, volume 2, 2007.

[12] Hyungsuk Ji, Sabine Ploux, and Eric Wehrli. Lexical knowledge representation with contextonyms. In *Proceedings of MT Summit IX, New Orleans, USA*. Association for Machine Translation in the Americas, 2003.

[13] B. Liu. Sentiment analysis and subjectivity. *Handbook of Natural Language Processing,*, pages 627–666, 2010.

[14] H. Liu and P. Singh. ConceptNet-a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4) :211–226, 2004.

[15] C. Ma, H. Prendinger, and M. Ishizuka. A chat system based on emotion estimation from text and embodied conversational messengers. *Entertainment Computing-ICEC 2005*, pages 535–538, 2005.

[16] R. Mihalcea and D. Radev. *Graph-based natural language processing and information retrieval*. Cambridge University Press, 2011.

[17] G.A. Miller. WordNet : a lexical database for English. *Communications of the ACM*, 38(11) :39–41, 1995.

[18] E. Pianta, L. Bentivogli, and C. Girardi. Developing an aligned multilingual database. In *Proc. 1st Int. Conference on Global WordNet*, 2002.

[19] Christopher Potts. Sentiment analysis tutorial, November 2011.

[20] D. Rao and D. Ravichandran. Semi-supervised polarity lexicon induction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 675–682. Association for Computational Linguistics, 2009.

[21] O. Serban, A. Pauchet, A. Rogozan, and J.-P. Pecuchet. DDMCE : recherche de cliques maximales dans des graphes dynamiques de grande taille. In *Proceedings of the Journée thématique : Fouille de grands graphes*. Réseaux : Approches Mathématiques et Informatique, 2012.

[22] P. Singh, T. Lin, E. Mueller, G. Lim, T. Perkins, and W. Li Zhu. Open mind common sense : Knowledge acquisition from the general public. *On the Move to Meaningful Internet Systems 2002 : CoopIS, DOA, and ODBASE*, pages 1223–1237, 2002.

[23] S. Stamou, K. Oflazer, K. Pala, D. Christoudoulakis, D. Cristea, D. Tufis, S. Koeva, G. Totkov, D. Dutoit, and M. Grigoriadou. Balkanet a multilingual semantic network for the balkan languages. In *Proceedings of the International Wordnet Conference, Mysore, India*, pages 21–25, 2002.

[24] P.J. Stone, D.C. Dunphy, and M.S. Smith. The general inquirer : A computer approach to content analysis. 1966.

[25] C. Strapparava and R. Mihalcea. Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied computing*, pages 1556–1560. ACM, 2008.

[26] C. Strapparava and A. Valitutti. WordNet-Affect : an affective extension of WordNet. In *Proceedings of LREC*, volume 4, pages 1083–1086. Citeseer, 2004.

[27] Y.R. Tausczik and J.W. Pennebaker. The psychological meaning of words : Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1) :24–54, 2010.

[28] A. Valitutti, C. Strapparava, and O. Stock. Lexical resources and semantic similarity for affective evaluative expressions generation. *Affective Computing and Intelligent Interaction*, pages 474–481, 2005.

[29] P. Vossen. *EuroWordNet : a multilingual database with lexical semantic networks*. Kluwer Academic, 1998.

[30] H.G. Wallbott, K.R. Scherer, et al. Emotion and economic development-Data and speculations concerning the relationship between economic factors and emotional experience. *European journal of social psychology*, 18(3) :267–273, 1988.

[31] J. Wiebe, T. Wilson, and C. Cardie. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2) :165–210, 2005.