Annotations for Opinion Mining Evaluation in the Industrial Context of the DOXA project

Patrick Paroubek[1], **Alexander Pak**[1], **Djamel Mostefa**[2]

[1]LIMSI-CNRS, BP 133
91403 Orsay cedex, France
{alexpak,pap}@limsi.fr

[2] Evaluation and Language resources Distribution Agency (ELDA)
55-57 rue Brillat-Savarin, 75013 Paris, France
mostefa@elda.org

Abstract

After presenting opinion and sentiment analysis state of the art and the DOXA project, we review the few evaluation campaigns that have dealt in the past with opinion mining. Then we present the two level opinion and sentiment model that we will use for evaluation in the DOXA project and the annotation interface we use for hand annotating a reference corpus. We then present the corpus which will be used on DOXA and report on the hand-annotation task on a corpus of comments on video games and the solution adopted to obtain a sufficient level of inter-annotator agreement.

1. Introduction

Along with an interest for incorporating emotions in technological devices, the recent years have seen the emergence of automatic opinion and sentiment analysis methods (B.Pang and L.Lee, 2008) particularly in the image management and survey business. Opinions are carried over various media, the press, web sites, radio, television etc. They are a spontaneous source of information, which is updated daily and provides the means to draw quickly an image of the perception that the public entertain with respect to some service, product or major actor of the entertainment or political scene. Survey and analysis of these information sources provide a company with a better knowledge of its customers. They give the means to anticipate new demands, to ensure their fidelity and to reduce attrition risks.
The DOXA[1] project aims at specifying and developing components, resources and services which will allow to :

- Automatically detect topics addressed in large volumes of texts in French and in English,
- Automatically detect feelings and opinions expressed within large volumes of texts in French and in English,
- Automatically detect relations between feelings and opinions expressed and the topics concerned by these feelings and opinions,
- Transform extracted information from texts into structured information to combine this new information with structured information, associated with texts and their authors, to deduct synthetized and exploitable knowledge, by using techniques of data analysis,
- Integrate the components of texts and data analysis into a new version of the INFOM@GIC's platform (services oriented) to build three applications dedicated to "opinion watch" , "consumers and citizens intelligence" , "customer loyalty and churn" for the end-users of the project : OpinionWay, EDF and Meetic.

The applications developed for end-users will help to survey in dynamic, quantitative and qualitative ways:

- the positioning of consumers, customers and users,
- the relationships they maintain with the universes about which they express themselves,
- the trends or evolutions of these universes.

They will help to improve both decision-making (On-Line Analytic Processing, segmentation, scoring, etc.) and operational processes (profiling), by integrating enriched knowledge into these processes.
In the next section, we will make a rapid survey of the various models we have found in the literature in relation with opinion analysis, and we will draw a picture of their relative positions based on the information dimensions that they consider, as far as it is possible to provide an integrated view based on their widely varying characteristics. This will serve us to locate in the landscape the model of (Y.Yannik-Mathieu, 1991) which was used as starting point for our opinion model in DOXA. Then we will have a second state of the art section, but this time devoted to a rendering of the evaluation activities for opinion mining. Once the background picture has been set we will see how both previous topics are addressed in the context of DOXA with first a presentation of opinion model that will be used for annotating the evaluation corpus and second, a presentation of the annotation guidelines and toolkit.

[1] DOXA is a project (DGE n^{o} 08-2-93-0888) supported by the numeric competitiveness center CAP DIGITAL of Île-de-France region which aims among other things at defining and implementing an OSA semantic model for opinion mining in an industrial context. See http://www.projet-doxa.fr

2. Opinion Mining and Sentiment Analysis (OSA) Models

OSA models vary greatly in their orientation. They may be either oriented toward discovering expression of opinion based on more or less rational considerations, judgments or appreciations, either oriented toward the modeling and representation of the expression of the sentiment/emotions that one entertains about an object or an issue. They vary also greatly in the number of dimensions that they use to represent opinion or sentiments and in the granularity of their semantics.

According to (A.Esuli and F.Sebastiani, 2006), opinion mining consists both in searching for the opinions or sentiments expressed in a document and in acquiring new methods to automatically perform such analysis. The authors mentioned three main activities of the field:

- A1 developing language resources for opinion mining, e.g. building a lexicon of subjective terms;
- A2 classifying text according to the expressions of opinion contained;
- A3 extracting from text opinion expressions, taking into account the relationship that links the expression of opinion (the words expressing the opinion) to the source (the author of the opinion statement) or to the target of the expression of opinion (the object the opinion is about) (S.-M.Kim and E.Hovy, 2006).

To build our synthetic view of the various models we will make use a set of general "features", each one broadly associated with a particular information dimension. The previous definition of the activities associated with opinion mining, provides us with the four main features that we will use in our description of the various models, namely:

1. the *opinion marker*, i.e. the language items expressing an opinion (A1 & A3),
2. the *opinion polarity*, the more or less positive impression felt when one reads the opinion expression (A2),
3. the *source*, the (possibly indirect) reference to the beholder of the opinion (A3),
4. and the *target*, the reference to the object/issue/person about which an opinion is expressed (A3).

Among the other features that we will use to organize our presentation of the various models for opinion mining, we have:

- the *intensity*, i.e. the relative strength of an expression,
- the *theme/topic*, whether the models makes use of a representation of the topic addressed, in the document containing an expression of opinion,
- the *information*, the more or less factual aspect of the opinion expression,
- the *engagement*, the relative implication that the opinion holder is supposed to have in supporting his opinion expression.

Listing the features sets of all the models we have encountered and putting them into relation yielded a graph that is too complex to be easily displayed because of the numerous links. So we decided to sort our presentation features according to an arbitrary order based on the intuitive importance one would accord to a given feature if it were missing from an opinion statement. In our mind, an opinion statement which would mention only the *intensity* of an opinion without giving any indication of its *polarity* should be considered less informative for opinion mining. As a result, we put *polarity* before *intensity* in our arbitrary ordering and following the same train of thought, we have afterward: the *target*, the *information*, the *engagement* and and the *source*. Putting the *source* last may seem strange, but very often the source is not explicitly mentioned in a document, since the source is the author. Then we sorted the different models of opinion, first according to the number of "features" they display and second according to the relative position of their features in our arbitrary ordering. For instance a model having only the attribute *polarity* would be judged more generic than a model which would have both *polarity* and *target*. With this considerations in mind, the twenty different models organize themselves into a quasi linear sort. From the most generic to the most specific model, we have identified six levels in the hierarchy of models in Figure 1. The first level of our hierarchy lists authors who have not proposed any attribute in particular, but have addressed the subject of opinion and sentiment in language. They are associated in our representation to the most generic (top) attribute *OSA model*. Level 2 shows authors who do not have any *polarity* in their model and level 3 those who did not address *Intensity*, and so on. Then we have used the same methodology at each level to refine our hierarchy. At level 1, we find the models of (R.Quirk et al., 1985), (J.Kamps et al., 2004) and (S.Berthard et al., 2004). They have defined other attributes of opinion expression, like *polarity, intensity, target, information* etc. (R.Quirk et al., 1985) have introduced the notion of *private state* which regroups all the expressions of subjectivity like emotions, opinions, attitudes, evaluations etc. This notion is also present in the model of (J.Wiebe et al., 2005), (B.Pang and L.Lee, 2008). The models of (K.Dave et al., 2003), (P.Turney, 2002), (A.Harb et al., 2008), (S.Somasundaran et al.,), (S.-M.Kim and E.Hovy, 2006) and (V.Stoyanov et al.,) are located at level 2. The models of (T.Mullen and N.Collier, 2004), (V.Stoyanov et al.,) and (H.Yu and V.Hatzivassiloglou, 2003) were considered more specific than those of level 2 because they stressed the importance of *target* and *source* for opinion mining. The work of (Y.Yannik-Mathieu, 1991) is characterized by a categorization of verbs expressing feelings. The model of (J.R.Martin and P.R.R.White, 2005) deals with evaluative aspects. The authors have mentioned three subtypes of evaluation, characterized by their respective attributes which are: *attitude*, *engagement* and *graduation*. *Attitude* refers to values returned by judgement from one or more sources and can be associated to emotional responses. Its three subtypes are: *judgement*, *affect* and *appreciation*. *Engagement* explicits the position, the implication of the source with respect to its expression of opinion. It is one of the main characteristics of subjectivity. (J.R.Martin and P.R.R.White, 2005) have introduced the concept of *graduation* which is further declined using *force* and *focus*. It expresses the strength of the opinion expression, so we merged this concept with *intensity* in our representation. Models of (Y.Choi et al., 2005) and (E.Riloff et al., 2003) are also at the same level as (E.Riloff et al., 2006), (P.Turney and M.Littman, 2003) and (Y.Yannik-Mathieu, 1991). At last, (B.Pang and L.Lee, 2008) and (J.Wiebe et al., 2005) propose the most complete models which regroup together all our presentation features.

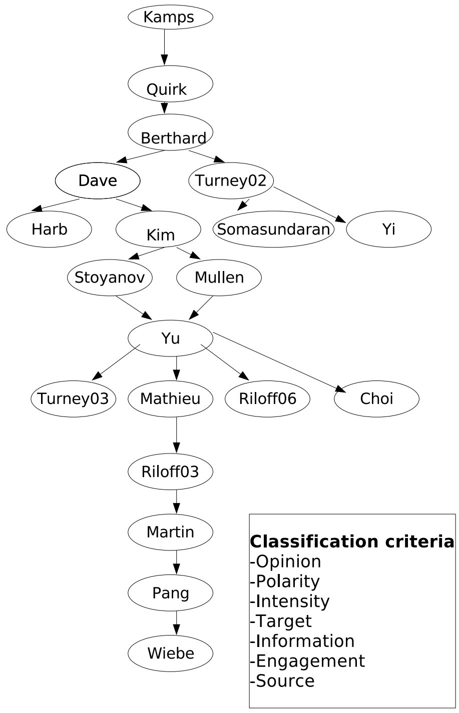


Figure 1: Hierarchical OSA models sorting from the most generic (top) to the most specific (bottom), based on the attribute ordering shown in the box, with *Opinion* being considered the most important and *Source* the least. When the same author has proposed different models, they are distinguished by publication year.

The characteristics that we retain from this survey as important for an OSA model are the three part representation with: the opinion/sentiment expression, its source and its target, and the distinction between sentiments, related to emotions from opinions which retain more a flavor of rationality. We will now survey the evaluation campaigns that proposed an evaluative framework to compare algorithms for OSA.

3. Opinion Mining Evaluation Campaigns

From 2006 up to 2008, TREC (Text Retrieval Conference) has proposed the *Blog Track* [2] for searching a corpus of blogs. The task consisted of:

1. distinguishing whether a post was subjective or objective (no opinion expressed in the document),
2. separating positive posts from negative ones,
3. sorting them by decreasing order of positiveness.

Performances observed during the campaign were measured in terms of Mean Average Precision (MAP), and the value measured ranged from 0,17 to 0,45.

Since 2006, the NTCIR-MOAT campaign (Multilingual Opinion Analysis Task) proposes to tag newspaper articles. Each sentence must be tagged depending on its relative subjectivity/objectivity and its relevance with respects to an *a priori* topic. At the clause level, the participants must identify the source and target of opinion, as well as the polarity expressed on a 3 value scale (positive, neuter, negative). In this campaign, identifying subjectivity and relevance have yielded better performance results (F-scores between 0.41 and 0.92) than detecting the source, the target and the polarity of an opinion expression (F-scores between 0 and 0.75).

In 2007, SemEval campaign (Semantic Evaluations) has proposed the task *Affective Text* to explore the connections between emotion and semantic lexicons. The participant had to annotate news headlines along two dimensions, first, according to the emotion expressed, chosen among a list of 6 basic *emotions: anger, disgust, fear, joy, sadness, surprise*; and second, according to the polarity (positive/negative) of the opinion expressed. The best recognized emotions were: *sadness* and *fear* with respective maximum F-scores of 0.30 and 0.20. *Anger, joy and surprise* have obtained a maximum F-score of 0.15, while *disgust* was the emotion the most badly recognized with a null F-score.

Finally in 2008, TAC campaign (Text Analysis Conference) had a *Question Answering* task with opinion questions. There were two types of questions: a first batch of questions about the opinion bearer (*Who supports what?*) and a second one about the opinion itself (*What are the critics about...Why do people like...*) Here the F-scores obtained for this two types of questions have ranged from 0.01 to 0.17. In TAC 2008, there was also a task for automatics opinion abstract generation taking as input the answers proposed for a set of questions.

In France two evaluation campaigns took place respectively in 2007 (C.Grouin et al., 2007) and 2009 [3], in the context of the series of *text mining challenges* (DEFT). The first campaign dealt with the assignment of a polarity value to a text bearing an opinion (review of books or of video games for instance), taken from a three value range (bad, average, good)(J.-B.Berthelin et al., 2008). The second campaign asked the participants to determine whether a text was subjective or objective as a whole and locally to identify the subjective portions of a text taken from a newspaper corpus made of both factual articles and more opinion oriented ones. For both campaigns, the reference data have been built automatically from the metadata present in each corpus. Results from the 2007 campaign were different depending on the corpus processed. The best results were obtained with the video game reviews (with F-scores values ranging from 0.46 to 0.78) while the corpus of scientific article review was more problematic for the participants (F-scores between 0.40 and 0.57).

[2] http://ir.dcs.gla.ac.uk/wiki/TREC-BLOG/

[3] http://deft09.limsi.fr

In 2009, the participants had slightly better success at the global analysis task (with F-scores between 0.66 and 0.92 for the French corpus) than at the task of subjective text spans identification (F-score ranging from 0.65 and 0.91). Note that for the later, the reference data had been produced also by combining the data produced by the participants with a majority vote algorithm.

4. DOXA Opinion models

4.1. DOXA full opinion model

In DOXA, the model that was selected to serve as an initial basis for opinion mining is the one proposed by (Y.Yannik-Mathieu, 1991), which relies on a study of the verbs used in French to express sentiments. The linguistic theory underpinning its model is the Lexique-Grammaire (M.Gross, 1968). In Yannick-Mathieu's work, the verbs expressing sentiment were divided in homogeneous semantic classes (for instance in one class you could find all the verbs expressing joy) further refined into classes based on common linguistic characteristics, mainly sub-categorization frames. Since some of the classes of the initial model did not fit well the purpose of DOXA and the number of classes was deemed to big to be manageable in the considered applicative context, a new OSA model was designed especially for DOXA jointly by Thales, IGM, ARISEM, LIMSI and LIP6. Taking into account the need of the end-users of the DOXA project, three levels were defined in the DOXA OSA model (see Figure 2):

- *macro*, which corresponds to the document level,
- *meso*, for the paragraph level, defined as a fixed-size text span in order to avoid any variability because of the type of text processed,
- *micro*, for the sentence level annotations, with a classical definition of sentence based on syntactic and typographic analysis.

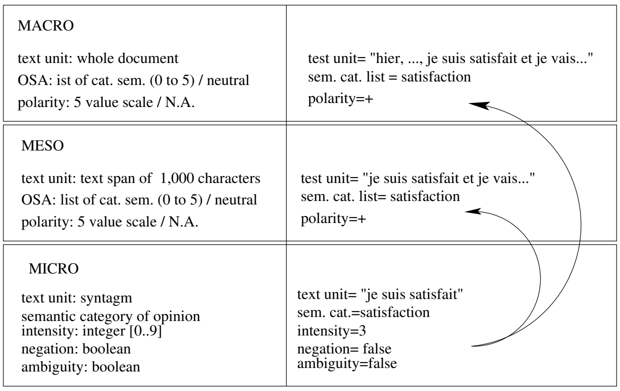

MACRO
text unit: whole document
OSA: ist of cat. sem. (0 to 5) / neutral
polarity: 5 value scale / N.A.
test unit= "hier, ..., je suis satisfait et je vais..."
sem. cat. list = satisfaction
polarity=+
MESO
text unit: text span of 1,000 characters
OSA: list of cat. sem. (0 to 5) / neutral
polarity: 5 value scale / N.A.
test unit= "je suis satisfait et je vais..."
sem. cat. list= satisfaction
polarity=+
MICRO
text unit: syntagm
semantic category of opinion
intensity: integer [0..9]
negation: boolean
ambiguity: boolean
text unit= "je suis satisfait"
sem. cat.=satisfaction
intensity=3
negation= false
ambiguity=false

Figure 2: DOXA opinion annotation full model.

The informations for the macro and meso level will be synthesized from the information of the micro level. Note that both for the macro and meso levels, several categories must be given when the polarity has value "mixed". There are 17 semantic categories of opinion, split into three broad classes regrouping respectively : the ones associated to affect, the ones associated to intellectual appreciation and the one pertaining to both previous classes.

negative	*positive*
affect	
displeasure unpleasant surprise sadness boredom contempt anger fear shame	pleasure pleasant surprise appeasement
affect and intellectual appreciation	
unsatisfaction	satisfaction
intellectual appreciation	
devalorisation agreement	valorisation disagreement

Table 1: The 17 semantic categories of opinion of the full DOXA model

4.2. DOXA evaluation opinion model

DOXA evaluation will address only the macro and meso levels because of cost and complexity reasons. The attributes and their values used in the evaluation opinion model are given in Table 2. In the model for evaluation we have a three value scale and an alternative value "neutral" for the the documents considered not to express any particular opinion, combined with a two value scale for intensity, which distinguishes weak-medium expression of opinion from strong ones. This representation, which is completely equivalent to the "neutral" plus five value scale of the full DOXA model, was preferred because deemed more intuitive for the human annotators. When "neutral" is used the polarity field is to be left undefined. Note also that the source of an opinion is not annotated since we assume that the source is the author of the document. As a consequence, reported opinions are ignored.

The idea behind the design of the DOXA evaluation opinion model is to have a model that is as much as possible similar to the DOXA full model and when not, to have a model that can be considered as generalization of the latter, so mapping from the full model to the evaluation model will be straightforward. Having two different models allows more freedom in the management of the project since the work on the evaluation procedure can start before the DOXA full model for analysis is completely finalized, and thus the deployment of the evaluation can happen sufficiently early in the project lifetime to be able to provide useful feedback on the technology developed. The *justification* feature of the DOXA evaluation model is used by the annotators to identify the text span of the document (or paragraph depending whether they annotate at the meso or macro leve) that represents best the opinion annotations (see example in Table 6. the *justification* and *text* information). This information will not be for performance scoring, but for improving the consistency checking during manual annotation and for en-

attribute	value
semantic category	recommandation_suggestion demand_query *a list of 1 to 5 DOXA semantic category* of opinion
polarity	$-, +/-, +$ neutral
intensity	weak-medium, strong
topic	*the target of the opinion expression* *taken from the current domain taxonomy* *(a list of 1 to 5 concepts)*
justification	*reference to the paragraph/text segment* *that represents best the opinion* *expressed in the document*
link	*when sereval cat. sem. and several topics* *associations linking specific pairs* *of topic/cat. sem. (see section eval. metrics)*

Table 2: DOXA macro/meso annotation of opinion.

riching the annotation guidelines by building lists of opinion related expressions.

5. Evaluation metrics

For evaluation, three metrics will be considered separatly: one for opinion annotations, one for topic annotation and one for the links between topic and opinion. In Table 3 we present the different metrics used. For topic annotation, we will use the conceptual distance proposed by Wu et Palmer (Z.Wu and M.Palmer, 1994). If we call $d(x, y)$ the topological conceptual similarity between two concepts of an ontology, i.e. the number of concepts encountered along the shortest path linking x to y, the conceptual similarity between two concepts C_1 and C_2, whose least common superconcept is the concept C_{ca} in a hierarchie of root R, is given by the formula : $Sim(C_1, C_2) = \frac{2\times d(C_{ca},R)}{d(C_1,C_{ca})+d(D_2,C_{ca})+2\times d(C_{ca},R)}$ The performance measure

Table 3: Evaluation metrics

annot.	evaluation function	value range
pol.	$equal(P_1, P_2)$	$\{0, 1\}$
cat.	$equal(C_1, C_2)$	$\{0, 1\}$
topic	$\frac{max_{k=1,m}\sum_{i=1}^{n} Sim(T_i,T_k)}{max(n,m)}$	$[0, 1]$
link	$\frac{max_{k=1,m}\sum_{i=1}^{n} Sim(T_i,T_k)\times equal(C_i,C_k)}{max(n,m)}$	$[0, 1]$

of a particular link annotation will be given by the product of the performance measure of the opinion annotation (0 or 1 value) by the performance measure of the topic (value obtained with Wu & Palmer conceptual similarity measure). The problem of how to compute the performance measure for the link information when the hypothesis data or the reference data contain several links will be solved by simple combinatorics, computing all possible performance values and keeping the highest performance value (minimal penalty for the system).

Parag id="d1009.1"	
Polarity:	positive
Intensity: *Text:*	strong *véritablement abouti* *[truly successful]*
Semantic Category: *Text:*	Interest_Valorization_Appreciation *véritablement abouti*
Topic: *Text:*	VideoGame *Arthur Et Les Minimoys*
Justification:	*Le studio français Étranges Libellules* *s'est lancé dans l'aventure pour* *nous proposer, au final, un titre* *véritablement abouti et soigné.* *[The French studio Strange Dragonflies* *embarked on the adventure to propose,* *in the end, a neat and truly successful title.]*

Table 4: Annotation example from the video game corpus with the DOXA evaluation opinion model.

6. Hand annotation tasks

The corpus is made of video games reviews, news and posts about video games collected from 8 dedicated web sites. There are 8,000 documents of an average size of 4500 words, split into fixed size paragraphs of 200 words (around 1000 characters). Right now there are around 2,000 documents which have been hand annotated with the DOXA OSA evaluation model (see details in Table 6).

As annotation software, we use the Knowtator[4] (P.Ogren, 2006) plugin for Protégé since the resulting software combination provides an annotation graphic interface coupled to an ontology browser for annotating opinion and sentiment in corpora (see Figure 3).

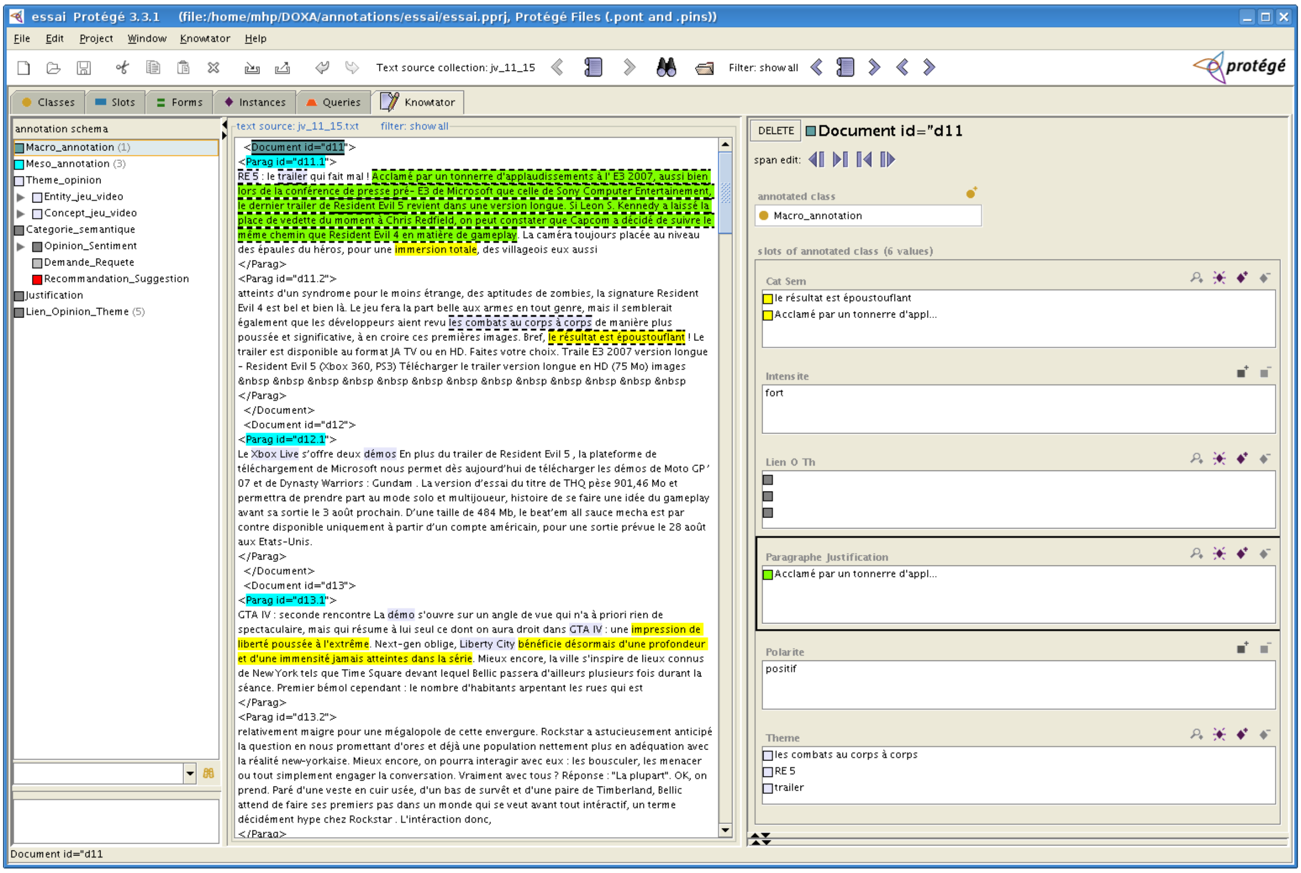
The annotation speed measured during this period, measured per day per annotator, ranged from 26 paragraphs (3 documents) to 41 paragraphs (5 documents).

The first return from the hand annotation task that started in october 2009 with two annotators are the low kappa values, in particular for documents which are only slighty positive or negative in nature and for the topic annotation. Table 6. displays the two measurements of kappa computed considering two different equality functions for topics and semantic category lists :

1. $equal(A, B) \equiv A \cap B \neq \emptyset$
2. $equal(A, B) \equiv \frac{|A\cap B|}{|A\cup B|} > 0.25$

This poor κ values can be partially explained by the fact that annotation guidelines were being finalized during this

[4] http://knowtator.sourceforge.net/


essai Protégé 3.3.1 (file:/home/mhp/DOXA/annotations/essai/essai.pprj, Protégé Files (.pont and .pins))
File Edit Project Window Knowtator Help
Text source collection: jv_11_15
Filter: show all
protégé
Classes
Slots
Forms
Instances
Queries
Knowtator
annotation schema
text source: jv_11_15.txt
filter: show all
DELETE
span edit:
annotated class
slots of annotated class (6 values)
Cat Sem
Intensite
Lien O Th
Paragraphe Justification
Polarite
Theme

- Macro_annotation (1)
- Meso_annotation (3)
- Theme_opinion
  - Entity_jeu_video
  - Concept_jeu_video
- Categorie_semantique
  - Opinion_Sentiment
  - Demande_Requete
  - Recommandation_Suggestion
- Justification
- Lien_Opinion_Theme (5)

Document id="d11

```
  <Document id="d11">
<Parag id="d11.1">
RE 5 : le trailer qui fait mal ! Acclamé par un tonnerre d'applaudissements à l' E3 2007, aussi bien
lors de la conférence de presse pré- E3 de Microsoft que celle de Sony Computer Entertainement,
le dernier trailer de Resident Evil 5 revient dans une version longue. Si Leon S. Kennedy a laissé la
place de vedette du moment à Chris Redfield, on peut constater que Capcom a décidé de suivre le
même chemin que Resident Evil 4 en matière de gameplay. La caméra toujours placée au niveau
des épaules du héros, pour une immersion totale, des villageois eux aussi
</Parag>
<Parag id="d11.2">
atteints d'un syndrome pour le moins étrange, des aptitudes de zombies, la signature Resident
Evil 4 est bel et bien là. Le jeu fera la part belle aux armes en tout genre, mais il semblerait
également que les développeurs aient revu les combats au corps à corps de manière plus
poussée et significative, à en croire ces premières images. Bref, le résultat est époustouflant ! Le
trailer est disponible au format JA TV ou en HD. Faites votre choix. Traile E3 2007 version longue
- Resident Evil 5 (Xbox 360, PS3) Télécharger le trailer version longue en HD (75 Mo) images
&nbsp &nbsp &nbsp &nbsp &nbsp &nbsp &nbsp &nbsp &nbsp &nbsp &nbsp &nbsp &nbsp
</Parag>
  </Document>
  <Document id="d12">
<Parag id="d12.1">
Le Xbox Live s'offre deux démos En plus du trailer de Resident Evil 5 , la plateforme de
téléchargement de Microsoft nous permet dès aujourd'hui de télécharger les démos de Moto GP'
07 et de Dynasty Warriors : Gundam . La version d'essai du titre de THQ pèse 901,46 Mo et
permettra de prendre part au mode solo et multijoueur, histoire de se faire une idée du gameplay
avant sa sortie le 3 août prochain. D'une taille de 484 Mb, le beat'em all sauce mecha est par
contre disponible uniquement à partir d'un compte américain, pour une sortie prévue le 28 août
aux Etats-Unis.
</Parag>
  </Document>
  <Document id="d13">
<Parag id="d13.1">
GTA IV : seconde rencontre La démo s'ouvre sur un angle de vue qui n'a à priori rien de
spectaculaire, mais qui résume à lui seul ce dont on aura droit dans GTA IV : une impression de
liberté poussée à l'extrême. Next-gen oblige, Liberty City bénéficie désormais d'une profondeur
et d'une immensité jamais atteintes dans la série. Mieux encore, la ville s'inspire de lieux connus
de New York tels que Time Square devant lequel Bellic passera d'ailleurs plusieurs fois durant la
séance. Premier bémol cependant : le nombre d'habitants arpentant les rues qui est
</Parag>
<Parag id="d13.2">
relativement maigre pour une mégalopole de cette envergure. Rockstar a astucieusement anticipé
la question en nous promettant d'ores et déjà une population nettement plus en adéquation avec
la réalité new-yorkaise. Mieux encore, on pourra interagir avec eux : les bousculer, les menacer
ou tout simplement engager la conversation. Vraiment avec tous ? Réponse : "La plupart". OK, on
prend. Paré d'une veste en cuir usée, d'un bas de survêt et d'une paire de Timberland, Bellic
attend de faire ses premiers pas dans un monde qui se veut avant tout intéractif, un terme
décidément hype chez Rockstar . L'intéraction donc,
</Parag>
```

Document id="d11

annotated class: Macro_annotation

slots of annotated class (6 values)

Cat Sem: le résultat est époustouflant
Acclamé par un tonnerre d'appl...

Intensite: fort

Lien O Th:

Paragraphe Justification: Acclamé par un tonnerre d'appl...

Polarite: positif

Theme: les combats au corps à corps
RE 5
trailer

Figure 3: Knowtator annotation interface for DOXA. The left hand side window displays the opinion annotation ontology, the middle contains window the text currently annotated with the selected annotation highlighted and the right hand side window shows the current annotation attribute values.

trial phase of annotation and that the annotators were new to the task. The topic annotation will have to be reconsidered, maybe by automatic generalization going up in the conceptual hierarchy of several levels and improving the annotation guidelines. An example of annotation guideline is provide in table 6..

κ scores with $\|A \cap B\| \neq \emptyset$		
Annotation	*macro level*	*meso level*
polarity	0.609	0.564
intensity	0.667	0.453
topic	0.381	0.524
sem. cat.	0.808	0.667
κ scores with $\frac{\|A \cap B\|}{\|A \cup B\|} > 0.25$		
Annotation	*macro level*	*meso level*
polarity	0.609	0.564
intensity	0.667	0.453
topic	0.230	0.333
sem. cat.	0.451	0.387

Table 5: Two measurements of kappa computed considering two different equality functions for topics and semantic category lists.

Number of documents	1,970
Number of paragraphs	18,415
Number of sentences	88,305
Number of words	202,345
Number of charcters	8,918,468
Characters per word	4.4

Table 6: Characteristics of the reference corpus for video games.

Polarity *Intensity*	*positive* *strong*
text characteristics	opinion expressions are very positive AND negative opinion expressions are missing or present in small number and very moderate in nature.
sem. cat.	only positive semantic categories should be present.

.

Table 7: Example of annotation guideline, for a positive polarity with a strong intensity.

7. conclusion

We have presented the context of the project DOXA which aims at developing an opinion mining industrial plateform. After reviewing the state of the art in terms of opinion mining modeling and opinion mining evaluation, we have presented the two opinion mining models developed in DOXA. One model will be used for fine grained opinion mining analysis and the other one, more general, will be used for evaluation. The coexistence of the two models within the same project finds its justification in project management considerations. Having two models enables to start early in the project lifetime the evaluation activities and thus we will be able to benefit from the evaluation feedback within the timeframe of the project. Then we have presented the video game reviews corpus we will use for evaluation along with the annotation toolkit and guidelines, as well a preliminary assessment of the inter-annotator agreement estimated during the initial start up phase of hand annotation. The preliminary results for the $kappa$ are relatively low, in particular for documents which are only slighty positive or negative in nature and for the topic annotation. To improve our $kappa$ for topic annotation, we will consider an automatic generalization going up in the conceptual hierarchy of several levels either during the manual annotation by providing feedback inside the annotation interface, or as post-processing of the annotated data.

8. Acknwoledgements

Thanks to Catherine Gouttas from THALES Communications (TCF) for helping us present the DOXA project.

9. References

A.Esuli and F.Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, pages 417–422, Genova, Italy.

A.Harb, M.Planitié, P.Poncelet, M.Roche, and F.Trousset. 2008. Détection d'opinions:apprenons les bons adjectifs. In *In Actes de l'Atelier Fouille des Données d'Opinions, conjointement Conférence INFORSID 08*, Fontainebleau, France, mai.

B.Pang and L.Lee. 2008. *Opinion mining and sentiment analysis*, volume 2. Now Publisher Inc, January.

C.Grouin, J.-B.Berthelin, S.El Ayari, T.Heitz, M.Hurault-Plantet, M.Jardino, Z.Khalis, and M.Lastes. 2007. Présentation de deft'07 (defi fouille de textes). In *Actes de l'atelier de clôture du 3ème DEfi Fouille de Textes*, pages 1–8, Grenoble, July. AFIA.

E.Riloff, J.Wiebe, and T.Wilson. 2003. Learning subjective noun using extraction pattern bootstrapping. In *Proceedings of the 7th Conference on Natural Language Learning*, pages 25–32, Edmonton, Canada.

E.Riloff, S.Patwardhan, and J.Wiebe. 2006. Feature subsumption for opinion analysis. In *In : Proceedings of EMNLP*.

H.Yu and V.Hatzivassiloglou. 2003. Towards answering opinion questions:separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of EMNLP*, pages 129–136, Sapporo, Japan.

J.-B.Berthelin, C.Grouin, M. Hurault-Plantet, and P.Paroubek. 2008. Human judgement as a parameter in evaluation campaigns. In *Proceedings of the Coling Workshop on Human Judgements in Computational Linguistics (HJCL 2008)*, Manchester, August.

J.Kamps, M.Marx, R.J.Mokken, and M.de Rijke. 2004. Using wordnet to measure semantic orientation of adjectives. In *Proceedings of LREC*, volume IV, pages 174–181.

J.R.Martin and P.R.R.White. 2005. *The Language of Evaluation:Appraisal in English*. Palgrave Macmillan, illustrated edition.

J.Wiebe, T.Wilson, and C.Cardie. 2005. Annotating expressions of opinions and emotions in language. Kluwer Academic Publishers, Netherlands.

K.Dave, S.Lawrence, and D.M.Pennock. 2003. Mining the peanut gallery:opinion extraction and semanctic classification of product reviews. In *Proceedings of the 12th International World Wide Web Conference*, Budapest, Hungary.

M.Gross. 1968. *Grammaire Transformationnelle du Français*. Larousse.

P.Ogren. 2006. Knowtator: A protégé plug-in for annotated corpus construction. In ACL, editor, *Proceedings of the Conference of the North American Chapter of the ACL on Human Language Technology: companion volume: demonstrations*, pages 273–275, New-York.

P.Turney and M.Littman. 2003. Measuring praise and criticism:inference of semantic orientation from association. pages 315–346. ACM Transactions on Information Systems.

P.Turney. 2002. Thumbs up or thumbs down?semantic orientation applied to unsupervised classification of reviews. In *Proceedings of 40th Meeting of the ACL*, pages 417–424.

R.Quirk, G.Leech, and J.Startvik. 1985. *A Comprehensive Grammar of the English Language*. NewYork:Longman.

S.-M.Kim and E.Hovy. 2006. Identifying and analyzing judgment opinions. In *Proceedings of the Joint HLT / NACACL Conference*.

S.Berthard, H.Yu, A.Thornton, V.Hativassiloglou, and D.Jurafsky. 2004. Automatic extraction of opinion propositions and their holders. In *Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text*.

S.Somasundaran, J.Ruppenhofer, and J.Wiebe. Discourse level opinion relations:an annotation study. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pages 129–137. ACL.

T.Mullen and N.Collier. 2004. Sentiment analysis using support vector machines diverse information sources. In *Proceedings EMNLP-04*.

V.Stoyanov, C.Cardie, D.Littman, and J.Wiebe. Evaluating an opinion annotation scheme using a nex multi-perspective question and answer corpus. In *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text*, Avril.

Y.Choi, C.Cardie, E.Riloff, and S.Patwardhan. 2005. Identifying sources of opinions with conditional random fieldsand extraction patterns. In *Proceedings of HLT/EMNLP*.

Y.Yannik-Mathieu. 1991. Sciences du langage. In *Les verbes de sentiment – De l'analyse linguistique au traitement automatique*. CNRS Editions.

Z.Wu and M.Palmer. 1994. Verbs semantics and lexical selection. In *In Proceedings of the 32^{nd} ACL*, pages pages 133–138, Las Cruces. New Mexico State University.