

Chaîne de traitement symbolique pour l'analyse d'opinion - l'analyseur d'opinions de Synapse Développement face à Twitter

Baptiste Chardon, Sophie Muller, Dominique Laurent, Camille Pradel, Patrick Séguéla¹
(1) Synapse Développement, 5 rue du Moulin Bayard, 31000 TOULOUSE
{baptiste.chardon, sophie.muller, dlaurent, camille.pradel, patrick.seguela}@synapse-fr.com

Résumé. Cet article présente la chaîne de traitement d'analyse d'opinion de Synapse Développement qui a été utilisée dans le cadre de la campagne DEFT 2015. Cette chaîne repose sur une analyse syntaxique et sémantique généraliste du français et sur des lexiques généralistes parfois associés à des lexiques spécialisés selon des besoins spécifiques de clients. Nous présenterons en première partie le contexte de cette campagne d'évaluation. La deuxième partie donnera une vue d'ensemble de la chaîne de détection de l'opinion proposée par Synapse. La troisième et la quatrième partie seront consacrées aux adaptations des ressources syntaxiques et lexicales au contexte du tweet et au lexique de spécialité relatif au développement durable. La partie 5 permettra de décrire la gestion des opérateurs et l'agrégation des opinions au niveau du tweet. Nous concluons sur les résultats obtenus en termes de classification selon la polarité sur les corpus d'apprentissage et de test de DEFT 2015.

Abstract.

Symbolic pipeline for opinion mining – Synapse opinion miner vs. Twitter.

This paper presents the Synapse Développement processing system analyzing opinions in discourse which has been used during DEFT 2015 campaign. This system is based on syntactic and semantic analysis and general lexicons sometimes associated with specialized lexicons according to specific needs of clients. In first part, we will present the context of this evaluation campaign. The second part will give an overview of the opinion detection system. The third and the fourth parts will be dedicated to adaptations of syntactic and lexical resources in the context of tweets and the lexicon of specialty related to sustainable development. Part 5 will describe the management of the operators and the opinion's aggregation at the tweet level. We conclude on the results obtained in terms of classification according to the polarity on the learning and test corpus of DEFT 2015.

Mots-clés : Analyse d'opinion, extraction d'opinion, analyse de sentiment, analyse de subjectivité, détection d'émotion

Keywords: Opinion extraction, opinion mining, sentiment analysis, subjectivity analysis, emotion detection

1 Introduction et contexte

Avec le développement des réseaux sociaux, spécialement de la plate-forme de microblogging Twitter, donner son avis sur Internet est aisé et fréquent. Par conséquent, tout organisme souhaitant obtenir un retour concernant son e-réputation, l'accueil reçu par un produit, le sentiment général sur un sujet de société, etc. a à sa disposition un flux d'informations colossal : Twitter propulse en effet actuellement plus de 500 millions de tweets chaque jour, et un million de sites intègrent des tweets, comme fil d'actualité ou comme lien avec une communauté.

L'analyse d'opinion est un domaine très actif dans la recherche en traitement des langues naturelles. Une partie importante des efforts de recherche s'intéresse à l'analyse de l'opinion dans les tweets : citons notamment les travaux de (Pak & Paroubek, 2010), (Koloumpsis et al., 2011), (Jiang et al., 2011), (Prinyanthan et al., 2012), ou pour le français de (Brun et Roux, 2014).

Dans ce contexte, de nombreuses sociétés sont également intéressées par les outils d'analyse automatique de l'opinion. Synapse Développement propose une chaîne de traitement professionnelle de l'opinion basée sur une analyse syntaxique et sémantique de pointe, et l'évaluation DEFT 2015 nous permet de confronter nos technologies avec un corpus annoté de tweets.

Afin de répondre au mieux aux besoins spécifiques de clients professionnels, notre chaîne de traitement repose sur des mécanismes symboliques. Ceci nous permet de réagir rapidement sur des demandes clients, et de maîtriser l'impact d'une modification liée à un besoin spécifique. La section suivante décrit plus en détail cette chaîne de traitement.

2 Vue d'ensemble de la chaîne d'analyse d'opinion TextAnalyst *by Synapse*

L'objet du produit TextAnalyst *by Synapse* est d'extraire les signaux faibles du bruit ambiant, d'identifier les informations pertinentes pour une activité, de produire des vues synthétiques et facilement appréhendables du domaine d'étude. Ce produit permet, entre autres applications de cette extraction d'informations, d'annoter les opinions et de les caractériser.

La chaîne d'analyse d'opinion est composée de plusieurs modules en cascade / pipeline. Le schéma suivant récapitule cette architecture :

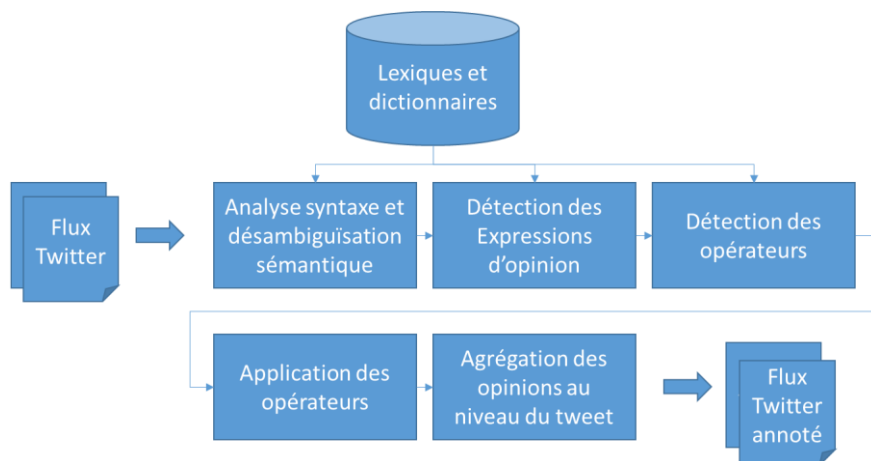


FIGURE 1. Aperçu de la chaîne de traitement.

Les modules utilisés sont les suivants :

- Connecteur d'entrée : permet la mise au format du texte d'entrée pour traitement par la chaîne d'analyse
- Analyse syntaxique et sémantique : il s'agit de l'analyseur syntaxique de Synapse Développement (cf. section 3)
- Reconnaissance des opinions unitaires : il s'agit d'un module de reconnaissance des expressions d'opinions en contexte. Ce module s'appuie sur l'analyse syntaxique et la désambiguïsation sémantique du module précédent, ainsi que sur des ressources lexicales, décrites en section 4.
- Reconnaissance et application des opérateurs : ce module détecte les opérateurs venant s'appliquer sur l'opinion (intensifieurs, négation, modalités) et applique leur effet sur les expressions d'opinion précédemment repérées. La section 5.1 décrit comment ces effets sont appliqués.
- Agrégateur d'opinion : ce module permet, à partir de la collection des expressions d'opinion repérées et d'éléments de contexte discursif, de calculer une opinion globale pour le document (ici le tweet). La section 5.2 décrit plus en détail le fonctionnement de ce module.
- Connecteur de sortie : permet l'écriture des sorties au format attendu pour l'évaluation DEFT

Les principales originalités de cette chaîne sont :

- le calcul de la modification (négation, modaux) de la polarité du segment via le modèle parabolique présenté dans (Chardon, 2013).
- la très large couverture du lexique généraliste (plus de 3000 entrées) créé avec le CNRS. Chaque entrée du lexique est associée à une catégorie d'opinion.
- l'aspect opérationnel de la chaîne de traitement, actuellement en production.

3 L'analyseur syntaxique et sémantique de Synapse

3.1 L'analyseur dans un contexte généraliste

Cordial (acronyme de CORrecteur D'Imprécisions et Analyseur Lexico-sémantique) est un analyseur syntaxique conçu à l'origine pour la correction orthographique et grammaticale. Développé au début des années 90, constamment maintenu et enrichi depuis, Cordial est le fondement de nombreux développements : composants linguistiques de nettoyage

automatique de texte, extracteur de mots-clés et de phrases-clés, extracteur de thèmes et de concepts, extracteur de terminologie et d'entités nommées, moteur de question-réponse.

Afin d'analyser le mieux possible un texte comportant éventuellement des fautes d'accord ou d'autres fautes, il est difficilement envisageable d'utiliser une grammaire formelle, qui fait le plus souvent appel à des règles d'appariement de groupes privilégiant le genre et le nombre. Nous avons donc associé à des règles générales et très peu liées aux attributs de genre et de nombre (analyseur à relâchement de contraintes) un ensemble d'outils statistiques, en particulier pour effectuer la désambiguïsation grammaticale.

Cordial est un analyseur à approche essentiellement statistique et probabiliste, même s'il utilise quelques règles pour la désambiguïsation grammaticale et surtout de très nombreuses règles pour la correction grammaticale.

Tous les mots sont alors "désambiguïsés" grammaticalement. En fait, la liste des formes grammaticales possibles pour chaque mot est conservée jusqu'à la fin de l'analyse, de même que la probabilité de chaque forme grammaticale, permettant des changements jusqu'à la phase de détection des relations entre les groupes. Un module spécifique traite alors les ambiguïtés "lourdes", c'est-à-dire celles constituées d'au moins deux mots ambigus, pour lesquelles les tables sont parfois insuffisantes et qui nécessitent la prise en compte d'un contexte supérieur à quatre mots. C'est le cas des couples adjectif/nom, nom/adjectif ("bonne alerte"), déterminant/nom ou personnel/verbe ("l'aide", "les avions").

Une première phase de désambiguïsation sémantique est alors effectuée. Elle se base essentiellement sur le contexte gauche et droit pour affecter une probabilité à chacun des sens possibles des mots polysémiques. Nous prenons en compte environ 25 000 sens pour 9 000 mots polysémiques, ce qui est un peu inférieur à un dictionnaire papier, le but étant de séparer des sens correspondant à des usages syntaxiques hétérogènes et surtout à des concepts nettement différenciés. Ainsi "abdomen" est monosémique pour Cordial alors que de nombreux dictionnaires distinguent la région inférieure du corps des mammifères et la partie postérieure du corps des arthropodes. Pour chaque mot, Synapse possède un nombre fini de sens (≤ 8). Ces sens, les informations qui les caractérisent ainsi que la façon de les différencier dans les textes, sont stockés dans le Lexique grammaire. Il référence 154 884 lemmes (noms, adjectifs, verbes, adverbes), 85 938 expressions nominales ("pomme de terre", "foie gras", etc.) et 12 059 expressions verbales ("avoir faim").

3.2 L'analyse syntaxique dans le contexte des tweets

Twitter est une plate-forme de microblogging qui permet de poster des messages courts, éventuellement liés à un contenu extérieur. Chaque message ne peut contenir que 140 caractères maximum, ce qui correspond généralement à une seule phrase. De ce format contraint découlent plusieurs spécificités :

- la texte est trop court pour donner un contexte au tweet. Pour pallier cela, les utilisateurs emploient fréquemment des « hashtags » ou « mots-dièse » permettant de spécifier de quoi il est question dans le tweet (e.g. : #Suisse, #ChangementClimatique, cf. figure N). Il peut également arriver qu'un hashtag ne soit pas neutre en termes d'opinion véhiculée (e.g. #PasCool). Ces hashtags, identifiés par le caractère '#', sont également utilisés par Twitter pour l'indexation des tweets dans leur moteur de recherche, et peuvent par la suite permettre à d'autres utilisateurs de retrouver des tweets sur un sujet donné.
- les tweets comportent fréquemment des liens vers des ressources externes sur le web, soit parce que le tweet porte sur le contenu d'un site web, soit parce que le site web contient un article détaillant le point de vue exprimé succinctement dans le tweet,
- les utilisateurs de Twitter sont identifiés par un pseudonyme unique. Ce pseudonyme ou nom d'utilisateur peut ensuite être utilisé dans un tweet pour référer à l'utilisateur en question. Ces noms d'utilisateurs sont précédés d'un '@'.



FIGURE 2. Deux tweets de @IGNFrance – captures d'écran de l'interface Twitter

La figure précédente montre deux exemples de tweets, publiés par le même compte @IGNFrance (compte officiel de l'Institut National de l'Information Géographique et Forestière – <https://twitter.com/IGNFrance>). On peut remarquer sur ces tweets qu'il y a deux manières d'insérer un de ces éléments dans le texte :

- le tweet du dessous, en date du 8 Mai, comporte plusieurs hashtags, une référence à un utilisateur, et une URL en fin de texte, non intégrés à la phrase.

- le tweet du dessus intègre directement le hashtag #ChangementClimatique dans la phrase, en remplacement du syntagme nominal « Changement climatique ».

Afin de préserver au mieux l'analyse syntaxique des tweets, l'analyseur syntaxique de Synapse Développement traite ces éléments de la façon suivante :

- les URL sont détectées et traitées comme dans n'importe quel texte

- les '#' et '@' sont ignorés, et les hashtags et nom d'utilisateurs sont traités comme des mots simples : si ceux-ci sont présents dans nos dictionnaires (e.g. : #changement #climatique), ils sont correctement reconnus et désambiguïsés. Si ceux-ci sont absents de nos dictionnaires, qu'ils représentent un syntagme (e.g. #ChangementClimatique), ou autre chose (e.g. #cop21), ils sont reconnus comme noms (commun ou propre) inconnu, et n'interfèrent donc pas avec la désambiguïsation sémantique du reste de la phrase.

4 Ressources lexicales pour l'analyse d'opinion

4.1 Ressources génériques issues de collaboration

Le lexique originel sur lequel se base notre chaîne de traitement est le lexique issu du projet CASOAR¹. Ce lexique de termes subjectifs se compose de 270 verbes, 632 adjectifs, 296 noms, 594 adverbes, 51 interjections et 178 expressions. Il a été construit manuellement, à partir de l'étude de corpus variés (articles de presse, commentaires web, et courrier des lecteurs). Chaque entrée possède une polarité et une intensité, ainsi qu'une catégorie sémantique, telle que définie par (Asher et al., 2008). Ces catégories sémantiques sont indépendantes d'une langue donnée.

Ce lexique gère à la fois l'ambiguïté de polarité et l'ambiguïté de sens : pour chaque entrée, seuls les sens subjectifs ont été associés à une polarité et une intensité. Ceci permet donc d'exploiter la désambiguïsation sémantique

Ainsi, lors de l'analyse syntaxique, si le sens détecté par l'analyseur n'est pas encodé dans le lexique, on peut en conclure que l'entrée lexicale a un sens objectif sinon elle est subjective. La figure suivante présente l'entrée "cher" : les balises "sense" définissent la catégorie sémantique, la polarité et la force de l'entrée lexicale, les balises "dicoSense" définissent un sens reconnu par la ressource dictionnaire associée. Une balise "sense" peut contenir une ou plusieurs balises "dicoSense".

```
<entry id="ADJ_cher">
  <lemma>cher</lemma>
  <pos>ADJ</pos>
  <sense type="jugement" category="évaluation" strength="1" polarity="pos">
    <dicoSense id="$1" description="aimé"/>
  </sense>
  <sense type="jugement" category="évaluation" strength="1" polarity="neg">
    <dicoSense id="$2" description="coûteux"/>
  </sense>
</entry>
```

Cette ressource est encore en cours de finalisation concernant le lien entre sens subjectif et sens des dictionnaires de Synapse Développement. À l'heure actuelle, seuls les liens pour les adjectifs sont complètement renseignés ; nous nous sommes donc limités à cette catégorie syntaxique pour les travaux présentés ici.

De plus, la chaîne dispose d'un « stop-lexique ». Ce lexique permet, dans des contextes spécialisés de supprimer l'annotation sur un sens, un lemme ou une expression. Par exemple, l'adjectif « écologique », porte généralement une connotation positive, mais est tout à fait contre-productif dans un corpus de tweets consacré au développement durable. Son annotation est donc désactivée par ce biais. De même, ce lexique permet de gérer des cas d'expressions figées qui ne correspondent pas vraiment à l'utilisation classique d'un terme. En effet, certaines expressions portent un sens si minoritaire qu'elles ne sont pas gérées dans le mécanisme de désambiguïsation sémantique. Ainsi le nom « top » est globalement positif dans son sens de supériorité, mais ne porte pas la même subjectivité dans l'expression « top départ ».

¹ projetcasoar.wordpress.com

De même, le terme « souci » est généralement associé à une polarité négative et à la catégorie Desinteret_Devalorisation_Depreciation mais ne sera pas porteur de cette notion lorsqu'il est employé dans l'expression « dans un souci de ».

Le lexique spécialisé contient donc des entrées du type :

```
<lex_entry amb_pol="false" amb_sense="false" id="SADV_dans_un_souci">
  <lemma> dans un souci
    <lemma_compose >
      <element lemma="dans" pos="1" rank="1"/>
      <element lemma="un" pos="2" rank="2"/>
      <element lemma="souci" pos="3" rank="3"/>
    </lemma_compose>
  </lemma>
  <pos>EXP</pos>
  <sense type="filter" category="Desinteret_Devalorisation_Depreciation" polarity="neutre"
strenght="0">
    <synapseSens id=" " description=" "/>
    <example/>
  </sense>
</lex_entry>
```

Enfin, nous avons mis en place un système de priorités sur les lexiques pris en compte dans le système. En effet, les termes absents de nos lexiques d'opinion génériques ainsi que les termes présents dans les bases Casoar et redéfinis en fonction du corpus de spécialité devaient permettre une annotation plus fine. De même, le lexique d'expressions complexes, en prenant plus en compte le contexte, permet un plus grand degré de certitude sur la qualification subjective qu'un terme isolé.

Ces constatations nous ont naturellement amenés à définir un ordre d'intervention de chaque lexique. Celui de spécialité arrivant naturellement en premier, suivi de celui des expressions complexes. Cette approche permet non seulement une meilleure précision mais évite aussi la multiplication de tags subjectifs sur une même expression, générant, par là même, des incohérences.

4.2 Extensions de ces ressources

4.2.1 Ressources lexicales à vocation généraliste

L'extension des ressources lexicales à vocation généraliste a été réalisée selon plusieurs méthodes. La confrontation à des textes de natures très différentes de ceux traités habituellement par notre système a nécessité l'élaboration d'une méthode pour augmenter le lexique spécialisé à partir du corpus d'apprentissage.

4.2.1.1 Méthode 1 : Méthode basée sur la coordination d'un adjectif avec un adjectif issu du lexique.

Les lemmes non porteurs d'une opinion d'après notre première analyse mais utilisés en coordination avec un terme porteur d'opinion sont des candidats à l'entrée dans le lexique.

- Les termes rencontrés avec des catégories différentes sont exclus.
- Les termes supprimés manuellement du premier lexique généré sont exclus.

L'intensité a été automatiquement fixée à 2, ce qui correspond à une intensité moyenne.

Cette méthode permet d'enrichir assez facilement le lexique mais n'apporte que peu d'amélioration en termes de rappel et de précision. En effet, l'ensemble des tweets dans lesquels ils entrent en coordination sont déjà bien classifiés. Ainsi dans le tweet numéro 48973538002144768, on relève la coordination entre « calme » et « reposant » qui propose « reposant » comme candidat à l'entrée dans le lexique avec les mêmes attributs de polarité et de catégorisation sémantique que « calme ».

```
Un endroit calme et reposant "les #étangs d'or". #beaune #cotedor #biodiversite #nature # vacances
http://t.co/c4sucgr2H1
```

Sur certains tweets, la coordination fait remonter des candidats plus difficiles à catégoriser, par exemple la coordination entre « saine » et « musclée » dans le tweet numéro 520630800509718528 :

Nouvelle recette saine et musclée sur le blog! </phr><phr id="phr_1_1g_38">Pas de gluten, 100% naturelle, 100% énergie de qualité! Réglez-vous! <http://t.co/XqHb5aBuSs>

Cette méthode n'a été appliquée que sur les adjectifs du fait de la plus grande variation de polarité observée entre deux noms ou verbes.

Pourtant certains cas peuvent se révéler intéressants, par exemple en ce qui concerne la coordination de « conseiller » et « assister » dans le tweet suivant, sachant que « conseiller » figurait déjà dans nos lexiques.

En effet, si cette méthode a validé l'intégration de « assister » suite à sa coordination avec « conseiller » dans des tweets tels que :

Steria conseille et assiste l'ESMA dans la gestion de ses projets IT et ses interactions avec son écosystème IT <http://t.co/mgwXUzfTzB>

4.2.1.2 **Méthode 2 : Méthode basée sur l'occurrence de certains lemmes dans des tweets portant la même polarité.**

Cette méthode a impliqué différentes sous-tâches afin de produire un lexique de qualité. L'idée était de pouvoir tester les ajouts selon diverses combinaisons de caractéristiques.

La première phase a consisté à analyser syntaxiquement et sémantiquement l'ensemble du corpus de tweets d'apprentissage. L'analyse syntaxique et sémantique a permis de relever les termes candidats les plus fréquents selon chacune de leur catégorie morphosyntaxique et selon chacun de leur sens. Dans un deuxième temps, le système d'apprentissage a attribué une polarité et une catégorie sémantique à chaque sens de chaque lemme.

A l'issue de cette phase, nous disposons d'un lexique de 20 976 termes candidats lemmatisés associés :

- à leur fréquence brute dans le corpus,
- à la fréquence d'apparition dans chaque groupe de tweets d'une même polarité
- et à la fréquence d'apparition dans chaque groupe de tweets d'une même classe sémantique.

Un premier filtrage des termes n'apparaissant que dans des tweets à polarité étiquetée neutre dans le corpus d'apprentissage a permis d'éliminer 7 987 lemmes candidats.

Objectif précision

Dans un premier temps, une analyse visant l'amélioration de la précision a été menée. Pour ce faire, nous avons conservé uniquement les lemmes qui n'apparaissent que dans des tweets portant la polarité positive ou négative. Ceci a permis de sélectionner 8 481 lemmes dont 7 737 n'apparaissent qu'une fois et 744 plus d'une fois dans le corpus selon les conditions suivantes :

- Un candidat retrouvé (hors négation) dans des passages à la fois étiquetés positivement et négativement est exclu (ex : ambiance) dans un premier temps
- Un candidat retrouvé (hors négation) dans des passages liés à des catégories sémantiques éloignées est exclu.

Un filtrage manuel de ces 744 lemmes a ramené cette liste à 239 lemmes.

Notons, que nous avons volontairement choisi de ne pas pousser plus avant l'apprentissage de termes et d'expressions et d'exercer un filtrage très strict afin de ne pas être victime d'un sur-apprentissage. Celui-ci aurait donné d'excellents résultats, tant sur le corpus d'apprentissage que sur celui de test, mais n'aurait revêtu que peu d'intérêt quant à l'évolution du système, voire l'aurait dégradé sur la plupart des corpus sur lesquels notre système pourrait intervenir. De fait, les lemmes intervenant moins de 3 fois dans le corpus d'apprentissage n'ont pas du tout été étudiés. Une analyse rapide des expressions relevées a donné les mêmes conclusions : peu d'entre-elles étaient utilisables sur corpus générique. La piste des expressions figées a donc été mise de côté.

À l'issue de ces apprentissages et après validation manuelle de ces candidats issus de la coordination et de la méthode d'analyse d'occurrence avec objectif précision, nous disposons d'un lexique comportant 51 adjectifs, 66 noms, 42 verbes et 6 adverbes. Soit, 167 lemmes désambiguïsés syntaxiquement et sémantiquement et étiquetés selon leur polarité, leur catégorie sémantique et leur intensité. Ils se répartissent comme suit : 68 lemmes/sens portant une polarité positive et 99 lemmes/sens portant une polarité négative

Objectif rappel

L'idée de cet objectif est de parvenir à repérer l'opinion portée par des termes qui apparaissent parfois en contexte positif, parfois en contexte négatif et parfois en neutre. Cette méthode permet de pallier les difficultés d'analyse des négations et de certains modaux liés à la structure même des tweets ainsi que les erreurs potentielles d'annotation dans le corpus d'apprentissage.

Le seuil d'analyse manuelle de ces candidats a été fixé à 80 % d'utilisation dans des tweets de même polarité afin de conserver la cohérence et de ne pas perdre des termes peu fréquents mais tout de même intéressants tels que « triste » qui apparaît dans 1 tweet étiqueté positif, 6 tweets étiquetés négatif et 1 tweet étiqueté neutre. Par contre, il sélectionne aussi quelques cas plus inattendus qui ne seront pas retenus ici. Par exemple, on ne s'attendrait pas à relever « transcanada » comme marqueur de subjectivité. Pourtant il apparaît dans 34 tweets négatifs contre 4 tweets positifs et 3 neutres.

De même on pouvait s'attendre à ce que les smileys, régulièrement considérés comme marqueurs objectifs de l'opinion, et même parfois comme éléments permettant l'évaluation d'un système, soient systématiquement dans des tweets portant la même polarité. Or il n'en est rien : ils sont régulièrement utilisés dans des tweets étiquetées par des polarités différentes, voire comme neutre.

Ainsi on relève les occurrences suivantes :

	tweets positifs	tweets négatifs	tweets neutres
: (1		
:- (1		1
:-))	1		
:)	7	4	11
:-)	4	1	6
:-/	1		
;-)	2	4	5
;-p		1	

TABLE 3 : Occurrences de smileys et polarité des tweets dans le corpus d'apprentissage

Ce filtrage a conduit à extraire 136 lemmes-candidats, revus manuellement.

À l'issue de cet apprentissage et après validation manuelle de ces candidats issus de la méthode d'analyse d'occurrence avec objectif rappel, nous disposons d'un lexique comportant 5 adjectifs, 17 noms, 20 verbes et 1 interjection. Soit, 43 lemmes désambiguïsés syntaxiquement et sémantiquement et étiquetés selon leur polarité, leur catégorie sémantique et leur intensité. Ils se répartissent comme suit : 15 lemmes/sens portant une polarité positive et 28 lemmes/sens portant une polarité négative

Les intensités ont été revues manuellement afin de conserver une cohérence avec les lexiques de la chaîne de traitement et de permettre une bonne gestion lors de l'agrégation des opinions au niveau du tweet.

Enfin, les termes ont été entrés un à un dans la chaîne de traitement afin de s'assurer une dernière fois de l'efficacité de ceux-ci.

4.2.2 Ressources spécifiques à Twitter : lexique de hashtags

Comme décrit dans la section 2, les tweets peuvent contenir un ou plusieurs hashtags, s'intégrant ou non dans la phrase. Dans le cas général, nos choix d'analyse syntaxique nous permettent de repérer les hashtags composés du dièse et d'un terme dans nos lexiques d'opinion.

Nous n'avons par contre pas souhaité étendre le lexique d'opinion à des termes n'ayant pas d'existence autre en français qu'en tant que hashtag (e.g., #Onsenfou). Pour cela, nous avons donc développé un lexique complémentaire permettant d'étiqueter comme positif certains hashtags. Ce lexique a été peuplé à partir d'une extraction des hashtags dans le corpus d'entraînement.

5 Agrégation des opinions : de l'expression d'opinion à la polarité d'un tweet

Une fois réalisée la résolution de la polarité au niveau du segment, un ensemble de règles permet de fusionner les opinions au niveau de la phrase ou d'un tag XML identifié dans le flux d'entrée.

Par défaut, 3 niveaux sont proposés : le niveau de la proposition, de la phrase et du document global. Dans le contexte de cette évaluation, le niveau choisi est donc le document global, c'est à dire le tweet, comme spécifié dans le guide d'annotation.

Afin de remonter les informations de polarité et d'intensité aux niveaux demandés, nous nous sommes appuyés sur le modèle de calcul d'opinion parabolique, développé dans le cadre de la thèse de Baptiste Chardon. Cette section décrit succinctement le modèle parabolique : pour plus de détails sur ce modèle, cf. (Chardon et al., 2013), (Chardon, 2013).

Le principe général derrière le modèle parabolique est une projection d'une opinion sur une parabole suivant sa polarité et son intensité. La figure suivante représente la parabole de projection.

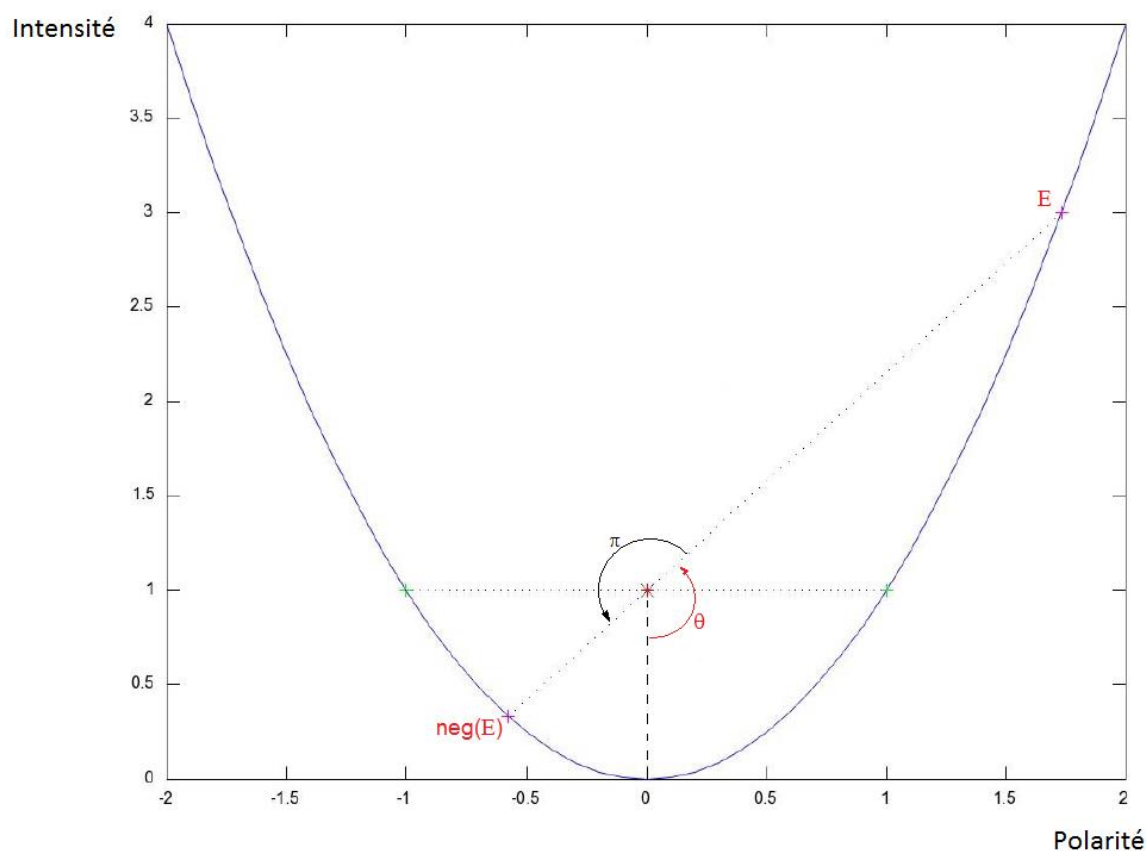


FIGURE 4. Exemple de parabole de projection

Le point E correspondant à une expression d'opinion est déterminé par son angle θ ; le signe de ce dernier est déterminé par la polarité de l'opinion, sa valeur absolue correspond à l'intensité de l'expression.

Par exemple, l'expression d'opinion "excellent" serait projetée au point E illustré sur la figure précédente.

Cette parabole associe un angle de $\pi/2$ aux opinions d'intensité standard, ce qui permet l'établissement de règles simples pour le calcul de l'effet des opérateurs.

5.1 Gestion des opérateurs

Notre lexique permet de reconnaître en contexte divers opérateurs s'appliquant sur l'opinion (négations, intensifications, opérateurs de modalité). Notre système prend en compte les opérateurs de négations et d'intensité, ainsi que l'effet de certaines modalités sur l'intensité de l'opinion.

5.1.1 Opérateurs de négation

L'opération de négation du modèle parabolique est basée sur deux hypothèses linguistiques, validées empiriquement dans les travaux (Chardon et al., 2013a).

- La négation renverse la polarité d'une expression d'opinion.
- La négation d'une expression d'intensité forte résulte en une expression d'intensité faible (par exemple, "excellent" étant d'intensité plutôt forte, "pas excellent" sera d'intensité plutôt faible).

Concrètement, l'opération consiste en un ajout de π à l'angle, comme illustré sur la FIGURE 4 par le point $\text{neg}(E)$.

Dans la littérature, le modèle de négation par décalage de [Taboada et al., 2010] ne satisfait pas à la première hypothèse proposée, tandis que le modèle de renversement simple (entre autres, [Choi et Cardie, 2008]) ne satisfait pas à la seconde.

5.1.2 Opérateurs d'intensification

L'opération d'intensification du modèle parabolique est implémentée comme un décalage à atténuation progressive : l'effet des opérateurs est maximal aux abords de l'intensité standard ($\pi/2$), et t lorsque l'on se rapproche des extrema, afin de maintenir une cohérence dans les valeurs retournées. Ceci revient à dire qu'intensifier des opinions déjà très intenses apporte moins de sens qu'intensifier une opinion d'intensité standard.

Concrètement, deux fonctions remplissant ces conditions sont :

$$\begin{cases} \text{int}_+(\theta) = \begin{cases} |\theta|/\theta * 2 * \mu * |\theta| & \text{si } |\theta| \leq \pi/3 \\ |\theta|/\theta * (\pi/2 + \frac{1}{2} * \mu * |\theta|) & \text{si } |\theta| > \pi/3 \end{cases} \\ \text{int}_-(\theta) = \pi - \text{int}_+(\pi - \theta) \end{cases}$$

avec μ un coefficient correspondant à la force de décalage de l'opérateur (donné par lexic, égal à 1 en l'absence d'information).

Cette opération se rapproche de l'implémentation simple de la littérature (décalage de +1/-1), avec une gestion plus fine des cas extrémaux.

5.2 Agrégation des opinions au niveau du tweet

Enfin, l'étape finale consiste en la fusion de l'ensemble des expressions d'opinions modifiées par les opérateurs en une information pertinente au niveau demandé. Pour cela, nous nous reposons sur deux mécanismes principaux : un ensemble de règles se basant sur une analyse surfacique du discours, et une heuristique de fusion.

Les règles de fusion discursive se basent sur les travaux de (Chardon et al., 2013b), qui décrivent un ensemble de règles permettant de fusionner deux opinions portées par deux segments de discours reliés par une relation SDRT (e.g. : conditionnelle, contraste). Dans le cas de notre chaîne de traitement, nous nous basons sur le repérage d'un nombre restreint d'opérateurs pour appliquer certaines de ces règles.

Si plusieurs opinions sont toujours dans le document une fois cette phase de traitement par règles effectuée, nous appliquons une heuristique numérique de fusion pour obtenir une polarité et une intensité au niveau du document. Cette heuristique favorise les opinions d'intensité tranchée dans le cas d'opinions portant la même polarité.

Dans le cas d'un document de la taille d'un tweet, repérer plusieurs expressions d'opinion dans le document est moins fréquent que sur un document non contraint. Sur l'ensemble des tweets récupérés en début de phase d'entraînement (7816), seuls 914 comportaient plusieurs expressions d'opinions repérées par la chaîne de traitement.

6 Résultats

Synapse Développement a participé à la campagne d'évaluation du 05 au 07 mai 2015.

6.1 Corpus d'apprentissage

La campagne d'évaluation DEFT 2015 propose de confronter les différents systèmes selon 3 tâches très distinctes :

- T1 : La classification des tweets selon leur polarité (positive, négative et neutre),
- T2 : La classification fine des tweets selon une classification de premier niveau comportant 4 classes génériques (information, opinion, sentiment et émotion) et 18 sous-classes liées au projet [uComp](#) (colère, peur, tristesse, dégoût, ennui, dérangement, déplaisir, surprise négative, apaisement, amour, plaisir, surprise positive, insatisfaction, satisfaction, accord, valorisation, désaccord, et dévalorisation).
- T3 : La détection de la source, de la cible et de l'expression d'opinion.

Pour notre part, nous avons choisi de n'évaluer notre système d'analyse de l'opinion que sur la première tâche de cette évaluation. En effet, ne disposant pas de la même classification sémantique des opinions, une correspondance n'aurait fait que dégrader les sorties du système. De plus, nous avons noté que l'intérêt principal des clients sur cette technologie réside essentiellement sur les notions de polarité. La catégorisation selon les classes sémantiques d'opinion est un résultat de recherche qui ne motive nos clients.

Notons qu'une analyse manuelle d'une sous-partie de ces tweets nous amène à penser que 15 à 20% de ces attributions de polarité sont, à tout le moins, discutables.

Prenons l'exemple du tweet identifié par le numéro 488730478061432832 :

Pluie de permis accordés par Ph. Henry co/ à Tinlot : 5 éoliennes entre Soheit et Terwagne <http://t.co/FGMyrPzGud> Pourquoi?

Ce tweet est classifié comme positif dans le corpus d'apprentissage, ce qui semble très douteux. De fait, les erreurs de ce type ont tendance à dégrader légèrement les résultats de la phase d'apprentissage. Les candidats à l'entrée au lexique de subjectivité de la chaîne de traitement étant revus manuellement, ils n'impactent pas la précision mais peuvent occasionner une légère baisse du rappel.

De même, certains tweets se répètent à l'identique dans le corpus et ne portent pas la même polarité. Une preuve de plus, s'il en fallait, de la grande difficulté d'attribution de l'opinion sur certains contenus, même par un humain. Il pourrait être intéressant de prendre en compte le taux d'accord inter-annotateurs lors de l'évaluation. Les erreurs sur les tweets à faible taux d'accord inter-annotateurs pourraient moins pénaliser un système que les tweets à polarité marquée et indiscutable.

Prenons l'exemple de la chaîne « Ségolène Royal: "Je ne veux pas que l'écologie soit punitive" », relevée pratiquement à l'identique dans 7 tweets, 6 annotés positivement, et 1 annoté négativement.

La chaîne initiale d'analyse d'opinion de Synapse, c'est-à-dire sans aucune adaptation au sujet abordé et au style de Twitter donnait un taux d'accuracy de 0,5658 et les résultats détaillés suivants :

Rappel <i>positif</i>	Précision <i>positif</i>	F-score <i>positif</i>	Rappel <i>négatif</i>	Précision <i>négatif</i>	F-score <i>négatif</i>	Rappel <i>neutre</i>	Précision <i>neutre</i>	F-score <i>neutre</i>
0,3111	0,5742	0,4036	0,3475	0,6666	0,4569	0,8565	0,5462	0,6671

TABLE 5 : Rappel, Précision et F-score avant adaptation au format tweet

Micro-précision	Macro-précision
0,564602661	0,595666667

TABLE 6 : Micro et Macro-précision avant adaptation au format tweet

Ces résultats, relativement décevants en baseline, s'expliquent du fait de la grande dépendance des annotations d'opinion, d'une part au contexte et aux relations de discours et d'autre part au sujet abordé.

L'intégration des principes d'adaptation à la tâche d'un point de vue de l'analyse syntaxique et de la prise en compte des spécificités du tweet (hashtag, smileys, patrons syntaxiques propres) ont permis d'atteindre un taux d'accuracy de 0,6632 et les résultats détaillés suivants :

Rappel <i>positif</i>	Précision <i>positif</i>	F-score <i>positif</i>	Rappel <i>néгатif</i>	Précision <i>néгатif</i>	F-score <i>néгатif</i>	Rappel <i>neutre</i>	Précision <i>neutre</i>	F-score <i>neutre</i>
0,4823	0,6752	0,5627	0,5407	0,7816	0,6392	0,8526	0,6272	0,7227

TABLE 7 : Rappel, Précision et F-score après adaptation au format tweet

Micro-précision	Macro-précision
0,657098567	0,694666667

TABLE 8 : Micro et Macro-précision après adaptation au format tweet

L'ajout du lexique d'adjectifs coordonnés et de termes issus de l'analyse d'occurrence en objectif précision a encore amélioré les retours du système pour atteindre un taux d'accuracy de 0,6725 et les résultats détaillés suivants :

Rappel <i>positif</i>	Précision <i>positif</i>	F-score <i>positif</i>	Rappel <i>néгатif</i>	Précision <i>néгатif</i>	F-score <i>néгатif</i>	Rappel <i>neutre</i>	Précision <i>neutre</i>	F-score <i>neutre</i>
0,4914	0,6817	0,5711	0,5683	0,7846	0,6592	0,8523	0,6371	0,7292

TABLE 9 : Rappel, Précision et F-score après ajout du lexique à visée précision

Micro-précision	Macro-précision
0,663961067	0,701133333

TABLE 10 : Micro et Macro-précision après ajout du lexique à visée précision

Enfin, l'ajout des termes issus de l'analyse d'occurrence en objectif rappel on amené le système à une accuracy de 0,6796 en fin d'apprentissage.

Rappel <i>positif</i>	Précision <i>positif</i>	F-score <i>positif</i>	Rappel <i>néгатif</i>	Précision <i>néгатif</i>	F-score <i>néгатif</i>	Rappel <i>neutre</i>	Précision <i>neutre</i>	F-score <i>neutre</i>
0,4901	0,6867	0,572	0,6066	0,7838	0,6839	0,8489	0,6448	0,733

TABLE 11 : Rappel, Précision et F-score après ajout du lexique à visée rappel

Micro-précision	Macro-précision
0,668538408	0,7051

TABLE 12 : Micro et Macro-précision après ajout du lexique à visée rappel

6.2 Corpus de test

Le corpus de test a été relevé le 6 mai 2015 et analysé et soumis le 7 mai 2015. Les résultats ont été reçus le 14 mai 2015 et une phase d'adjudication a suivi, jusqu'au 18 mai 2015. Les résultats suite à cette adjudication ne sont pas en notre possession lors de la rédaction de cet article. Soulignons toutefois, que sur les 1088 différences entre l'évaluation et les sorties de l'analyse d'opinion de la chaîne TextAnalyst ^{by Synapse}, nous relevons 448 attributions de polarité contestables (accord inter-annotateurs 100% pour 2 annotateurs internes). Notons aussi que certains tweets sont redondants, tout comme dans le corpus d'apprentissage, ce qui impacte d'autant plus lorsque ceux-ci sont d'une polarité qui peut prêter à débat.

La synthèse des résultats de l'ensemble des participants (12 équipes ont proposé des soumissions, pour 22 équipes inscrites) est présentée dans le tableau suivant :

Moyenne	Médiane	Ecart-type	Min	Max
0.5819492417	0.6933297846	0.238071241	0.0408344543	0.7359865456

TABLE 13 : Présentation de résultats de l'ensemble des participants

Les résultats par classe de polarité de la chaîne d'extraction de Synapse :

- Polarité négative : 0.767255216693419

Vrai positifs : 478 Faux positifs : 145

- Polarité neutre : 0.647663071391885
Vrai positifs : 1 261 Faux positifs : 686

- Polarité positive: 0.687268232385661
Vrai positifs : 556 Faux positifs : 253

Micro-précision	Macro-précision
0.679195028114827	0.700728840156988

TABLE 14 : Micro et Macro-précision lors de la phase de test

Références

BENAMARA, F., CHARDON, B., MATHIEU, Y., POPESCU, V., ASHER, N. (2012). How do negation and modality impact on opinions?. Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics- Association for Computational Linguistics, 10-18.

BRUN, C., ROUX, C. (2014). Décomposition des «hash tags» pour l'amélioration de la classification en polarité des «tweets». Actes de TALN. 146.

CHARDON, B. (2013). Chaîne de traitement pour une approche discursive de l'analyse d'opinion (Doctoral dissertation, Université de Toulouse, Université Toulouse III-Paul Sabatier)

CHARDON, B., BENAMARA, F., MATHIEU, Y., POPESCU, V., ASHER, N. (2013). Measuring the effect of discourse structure on sentiment analysis. Actes de *Computational Linguistics and Intelligent Text Processing*, 25-37. Springer Berlin Heidelberg.

CHARDON, B., BENAMARA, F., MATHIEU, Y., POPESCU, V., ASHER, N. (2013). Sentiment composition using a parabolic model. Proceedings of the *10th International Conference on Computational Semantics (IWCS 2013)*. 47-58.

CHARDON, B., BENAMARA, F., POPESCU, V. (2012). Projet CASOAR : le discours pour l'analyse de l'opinion. Actes des *Journée ATALA Discours et TAL : des modèles linguistiques aux applications*.

CHOI, Y., CARDIE, C. (2008). Learning with compositional semantics as structural inference for subsentential sentiment analysis. Actes de *Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. 793-801.

GODARD, D. (2013). Les négateurs. La grande Grammaire du français. Éditions Actes Sud

JIANG, L., YU, M., ZHOU, M., LIU, X., ZHAO, T. (2011). Target-dependent twitter sentiment classification. Proceedings of the *49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1* (pp. 151-160).

KOULOUMPIS, E., WILSON, T., MOORE, J. (2011). Twitter sentiment analysis: The good the bad and the omg!. *ICWSM*, 11, 538-541.

LAURENT, D., NÈGRE, S., SÉGUÉLA, P. (2009). L'analyseur syntaxique Cordial dans Passage. Actes de *TALN*, 9.

PAK, A., PAROUBEK, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *LREC*. Vol. 10, 1320-1326.

PRIYANTHAN, P., GOKULAKRISHNAN, B., RAGAVAN, T., PRASATH, N., PERERA, A. S. (2012). Opinion mining and sentiment analysis on a twitter data stream. *ICTer 2012*.

TABOADA, M., VOLL, K., BROOKE, J. (2008). Extracting sentiment as a function of discourse structure and topicality. Simon Fraser University School of Computing Science Technical Report.

ZHOU, L., LI, B., GAO, W., WEI, Z., WONG, K. (2011). Unsupervised discovery of discourse relations for eliminating intra-sentence polarity ambiguities. Actes de *Conference on Empirical Methods in Natural Language Processing (EMNLP'11)*.