

Rapport de Projet : Traitement Automatique du Texte en IA

Sujet :

Identification des opinions exprimées dans les avis et
commentaires d'un ordinateur.

Yann MARTIN D'ESCRIGNE
Yohann TOGNETTI

Table des matières

1	Introduction	2
1.1	Présentation générale du projet	2
1.2	Choix du sujet	2
2	Description des tâches	2
2.1	Entités	2
2.2	Catégories	2
2.3	Polarité	3
3	Jeux de données	3
3.1	Structure des jeux de données	3
3.2	Modification du jeu de données	3

1 Introduction

1.1 Présentation générale du projet

Dans le cadre de notre cours de Traitement automatique du texte en IA, il nous a été demandé d'effectuer un projet de notre choix impliquant une intelligence artificielle travaillant sur des données textuelles. La restriction sur le sujet se limitant à l'existence d'un jeu de données conséquent.

1.2 Choix du sujet

Notre choix de sujet fut **l'identification des opinions exprimées dans les avis et commentaires d'un ordinateur**. Plus précisément, comment retrouver parmi un ensemble de commentaires les parties de l'ordinateur visées, sur quels aspects et quel en est l'avis général qui en ressort.

2 Description des tâches

Ce sujet est fortement inspiré du "SemEval-2015 Task 12" sur la partie "Aspect Based Sentiment Analysis (ABSA) : Laptop Reviews" et partage donc le même objectifs avec des simplifications. Les différentes entités et catégories qui vont suivre proviennent également des annotations du sujet de SemEval. Chacun des commentaires est fragmentés phrase par phrase afin de faciliter le travail de l'IA.

2.1 Entités

Tout d'abord il s'agit de retrouver dans la phrase le(s) entité(s) ciblée. Elles peuvent être l'ordinateur comme un tout, ses parties physique (clavier, écran..) , des logiciels ou OS (Windows, navigateur, jeux...) ou bien même une compagnie et ses services. (DELL, Apple, le support technique, la livraison..).

Remarque : Rien n'interdit d'avoir plusieurs entités dans la même phrase.

Voici la liste des entités possible :

- DISPLAY (=moniteur, écran),
- CPU (=processeur),
- MOTHERBOARD (=carte mère),
- HARDDISC (=disque dur),
- MEMORY (=mémoire, RAM),
- BATTERY (=batterie),
- POWER_SUPPLY (=chargeur, unité de chargement, cordon d'alimentation, (power) adaptateur),
- KEYBOARD (=touche, clavier, pavé numérique),
- MOUSE (=souris, pavé tactile)
- FANS_COOLING (=ventilateur, système de refroidissement),
- OPTICAL_DRIVES (=lecteur CD, DVD ou Blue-ray),
- PORTS (=USB, HDMI, VGA, lecteur de carte),
- GRAPHICS (=carte graphique, carte video),
- MULTIMEDIA_DEVICES (=son, audio, microphone, webcam, haut-parleur, casque, écouteurs).

2.2 Catégories

Une fois obtenue il faut également trouver sur quelle(s) catégorie(s) de l'entité le commentaire porte. Cela peut être un aspect général , sa prise en main, ses performances, son design et ses fonctionnalités, etc...

Remarque : Là encore, rien n'interdit d'aborder plusieurs catégorie pour la même entité au sein d'une même phrase.

Voici la liste des catégories possible :

- GENERAL,
- PRICE (=prix),
- QUALITY (=qualité),
- DESIGN_FEATURES (=design et fonctionnalités),
- OPERATION_PERFORMANCE,
- USABILITY (=ergonomie, prise en main),
- PORTABILITY (=portabilité),
- CONNECTIVITY (=connectivité),

— MISCELLANEOUS (=divers).

L'entité E et la catégorie C qui s'y rapporte forme ainsi un couple E#C.

Remarque : Il est à noter la possibilité qu'aucun couple E#C ne se rapporte à une phrase d'un commentaire.

2.3 Polarité

Enfin, chaque phrase (et non chaque couple E#C comme dans le sujet de SemEval) se verra attribuer une polarité. Les valeurs de cette polarité sont : **negative** pour les phrases soulignant des défauts ou de mauvais avis, **positive** pour celles qui au contraire mettent en valeur des points ou des avis positifs, **neutral** lorsque la phrase ne met pas d'opinion en avant et donne par exemple un conseil et finalement **mixed** pour les textes critiquant un aspect de l'ordinateur mais appréciant un autre.

3 Jeux de données

Les jeux de données se divisent en deux groupes : le jeu de données d'entrée correspondant à ce que l'IA va utiliser pour s'entraîner et celui de test sur lequel les mesures seront effectuées. Tout deux proviennent du site du SemEval.

3.1 Structure des jeux de données

Ce sont tout deux des documents XML possédant la structure suivante (simplifiée) :

```
</sentence>
<sentence id="198:315">
  <text>The keyboard is good and easy to touch type on.</text>
  <Opinions>
    <Opinion category="KEYBOARD#GENERAL" polarity="positive"/>
    <Opinion category="KEYBOARD#USABILITY" polarity="positive"/>
  </Opinions>
</sentence>
<sentence id="198:316">
```

FIGURE 1 – structure du fichier XML

3.2 Modification des jeux de données

Suite à notre implémentation de notre IA qui sera décrite dans le chapitre suivant, de nombreux problèmes nous ont forcés à ajouter nous même des phrases dans le jeu de données d'entraînement.

En effet notre IA répond de la présence ou non de chaque couple E#C dans la phrase actuelle pour chaque combinaison d'entité E et catégorie C. Cela se traduit par une nécessité de nombreuses données appartenant à chacun des couples pouvant apparaître afin d'obtenir une fonction d'évaluation optimale.

Environ 350 phrases ont été ajoutées, toutes sont la section 'sentences' d'ID 198. Certaines proviennent d'avis de consommateur sur les ordinateurs portables les plus commentés sur Amazon (en anglais), notamment pour les phrases portant sur le clavier, écran, pavé tactile et batterie. Nous avons rédigés le reste pour les catégories plus complexes et moins abordées en générale. Par exemple la qualité des lecteurs DVD ou bien l'ergonomie des OS.

Remarque : Il est à noter que certains couples n'apparaissent ni dans les données d'entraînement, ni dans les données de test. Par simplification, ces couples ont été ignorés et aucunes phrases n'est étiquetées avec.