CANÉPHORE : un corpus français pour la fouille d'opinion ciblée

J. Lark
LINA - UMR CNRS 6241, Dictanova
joseph@dictanova.com

E. Morin
LINA - UMR CNRS 6241
emmanuel.morin@univ-nantes.fr

S. Peña Saldarriaga
Dictanova
sebastian@dictanova.com

Objectifs

- Fournir un corpus annoté réutilisable par la communauté
- Évaluer la performance d'un système de fouille d'opinion au niveau :
  → des sujets discutés
  → des aspects de ces sujets
  → des marqueurs d'opinion
- Évaluer la robustesse de l'analyse face à un corpus en français dégradé
- Expérimenter des méthodes d'extraction de marqueurs d'opinion

Fouille d'opinion ciblée

La fouille d'opinion ciblée (*aspect-based sentiment analysis*) consiste à extraire au sein de chaque phrase les **éléments clés parmi les arguments avancés** par des auteurs de critiques ou de commentaires.

Il s'agit par exemple de retrouver dans la phrase :
"*La salle du restaurant est bruyante*"
le sujet "'*restaurant*", l'aspect "*salle*" et le marqueur d'opinion "*bruyante*".

L'analyse à ce niveau de granularité permet d'**exploiter toute la richesse des prises de parole**, contrairement à celle au niveau du document entier. L'information extraite reflète donc plus fidèlement les avis ou arguments des internautes.

C'est particulièrement vrai lorsque qu'un auteur partage son opinion sur plusieurs sujets dans un même texte, ou sur plusieurs aspects d'un même sujet.

Cependant les corpus annotés permettant l'évaluation de cette tâche sont rares.
Nous présentons ici **un corpus français librement accessible** de tweets manuellement annotés. Nous accompagnons ce corpus de résultats préliminaires en extraction peu supervisée des marqueurs d'opinion annotés.

Le corpus est disponible sur la plate-forme GitHub. Les tweets doivent être téléchargés à partir d'une liste d' identifiants.

Description du corpus annoté

Documents

Le corpus est un ensemble de **10 000 tweets** échangés pendant l'événement "Miss France" en 2012 (soit environ **127 000 mots**). Les doublons, les retweets, les citations et les tweets considérés trop courts ont été retirés.

Les tweets présentent régulièrement des **erreurs orthographiques ou typographiques**, et un **vocabulaire imagé ou argotique**. Les principaux sujets du corpus sont les candidates du concours, dont les prestations sont jugées par les internautes.

Éléments annotés du corpus	**#**
Tweets contenant une annotation	5372
Sujets	708
Marqueurs	1967
Aspects	292
Triplets cible-aspect-marqueur	955
Couples cible-marqueur	5220

Annotation

L'annotation du corpus a été réalisée grâce à l'outil libre Brat, avec lequel il est possible d'étiqueter pour chaque tweet les **sujets, aspects ou mots marqueurs d'opinion** présents ainsi que leurs **relations**.

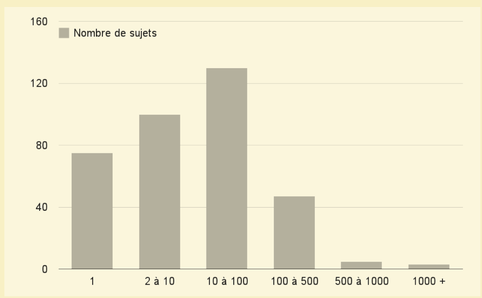
Pour chaque tweet exprimant une **opinion directe** sur un sujet explicitement mentionné, nous relevons les segments de texte les plus courts permettant à un humain sans connaissance du sujet de disposer d'une **information non ambigüe**.

Aucune annotation n'est effectuée dans les cas où :
- l'énoncé ne décrit pas *clairement* une opinion
- le sujet est uniquement désigné par un pronom
- l'expression de l'opinion dépend de connaissances *trop spécifiques*

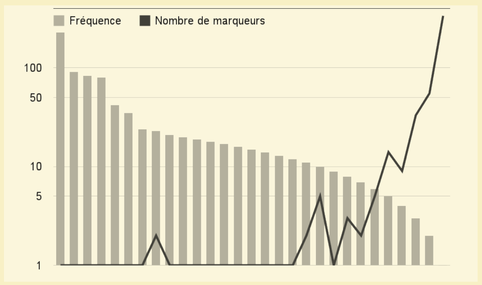
La nature **subjective** de ces considérations est cependant une source de difficulté pour l'étape d'annotation, puisqu'une opinion peut être exprimée au moyen d'une **construction complexe** ou de **connaissances implicites**.

Nous retrouvons ainsi un grand nombre de phrases exprimant une comparaison, soit entre deux sujets du corpus, soit impliquant une entité externe.

Marqueurs d'opinion	**#**
Positifs	687
Négatifs	1290
Mots (unigrammes)	1075
Dont uniques	740
Expressions (n-grammes, n > 1)	892
Dont uniques	800



Nombre de sujets uniques par fréquence



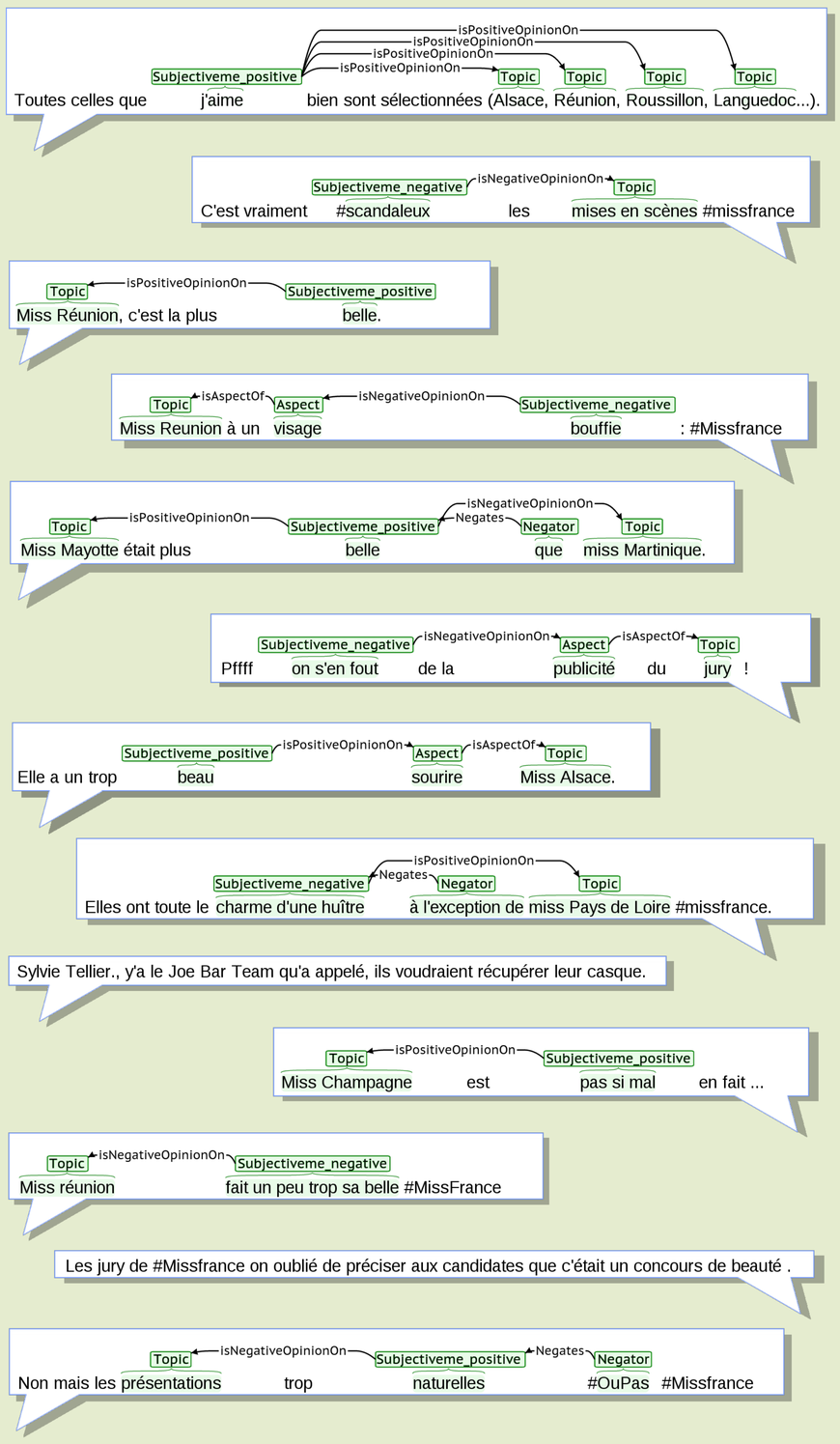
Nombre de marqueurs d'opinion uniques par fréquence

Utilisation du corpus

Le format d'annotation choisi permet d'évaluer la recherche de **chacun des éléments clés de la fouille d'opinion** : les sujets et aspects discutés ainsi que les mots marqueurs d'opinion.

La détection peut cependant être adaptée dans la mesure où une correspondance exacte des segments de textes choisis peut être trop restrictive.
Le corpus est disponible sous la licence GNU GPL V2.

Exemples d'annotations



Extraction des marqueurs d'opinion

Dans le cadre de ce travail nous nous intéressons en particulier à l'extraction des marqueurs d'opinion du corpus.

Nous expérimentons deux méthodes peu supervisées pour l'extraction des mots porteurs d'opinion, sans tenir compte de leur polarité.

Patrons fixes La première méthode consiste à identifier les patrons syntaxiques fixes les plus fréquents entourant un marqueur connu. Nous retenons les patrons syntaxiques dans une fenêtre de 7 mots au maximum.

Modèle SVM La seconde méthode consiste à réaliser un apprentissage sur les contextes morpho-syntaxiques des mots entourant un mot candidat dans une phrase. Les traits utilisés sont les étiquettes grammaticales et les lemmes des mots voisins.

Sujet	Patrons fixes			Modèle SVM		
	Précision	Rappel	F1	Précision	Rappel	F1
Miss Alsace	**96,47**	20,76	34,17	83,08	13,67	23,48
Miss Bretagne	**80,95**	8,76	15,81	80,36	23,20	36,00
Miss Réunion	**98,08**	28,02	43,59	96,77	16,48	28,17
Miss Languedoc	**94,59**	51,85	66,99	85,19	17,04	28,40
Miss Martinique	92,31	13,64	23,76	**95,45**	23,86	38,18
Miss Guyane	**95,00**	37,25	53,52	77,27	16,67	27,42
Miss Provence	**88,89**	11,27	20,00	83,33	7,04	12,99

Évaluation de l'extraction par occurrence des marqueurs

- L' extraction est amorcée à partir d'un lexique dont les mots ont été choisis pour leur faible variation de sens et de polarité.
- L' union des marqueurs obtenus par l'une et l'autre des deux méthodes proposées permet d'améliorer le rappel.

Perspectives

Nous espérons que la diffusion de ce corpus encouragera la création de ressources suivant le même type d'annotation.

Il serait intéressant de disposer d'un ensemble de corpus de natures différentes permettant d'évaluer la robustesse d'un système de fouille d'opinion.

Nos futurs travaux en extraction de marqueurs d'opinion seront orientés vers l'inférence de la polarité des marqueurs, et de leur expansion.

Nous envisageons pour cela de nous appuyer davantage sur les constructions phrastiques caractéristiques de la présence d'une opinion, ainsi que sur un plus grand nombre d'indices stables.