

Vers des pratiques collaboratives pour les systèmes d'organisation de connaissances

Amel Fraisse **Patrick Paroubek**

GÉRICO



amel.fraisse@univ-lille3.fr, pap@limsi.fr

10ème Colloque ISKO-France 2015, 5-6 novembre 2015,
Collège Doctoral Européen, Strasbourg

Plan

Introduction

Les systèmes d'organisation de connaissances

Le crowdsourcing pour la construction collaborative des connaissances

Notre approche pour la construction collaborative de connaissances factuelles et affectives

Conclusion

Plan

Introduction

Les systèmes d'organisation de connaissances

Le crowdsourcing pour la construction collaborative des connaissances

Notre approche pour la construction collaborative de connaissances factuelles et affectives

Conclusion

Contexte

- ▶ les **Systemes d'organisation de connaissances** suscitent de plus en plus de travaux dans le domaine des **humanités numériques**
- ▶ Avant le Web 2.0 : les SOC sont conçus par des spécialistes de la connaissance
- ▶ Avec le Web 2.0 : différents types d'intervenants pour la construction de la connaissance
- ▶ **l'utilisateur final est devenu producteur de connaissance : les mots-dièses, les tags, etc.**

Plan

Introduction

Les systèmes d'organisation de connaissances

Le crowdsourcing pour la construction collaborative des connaissances

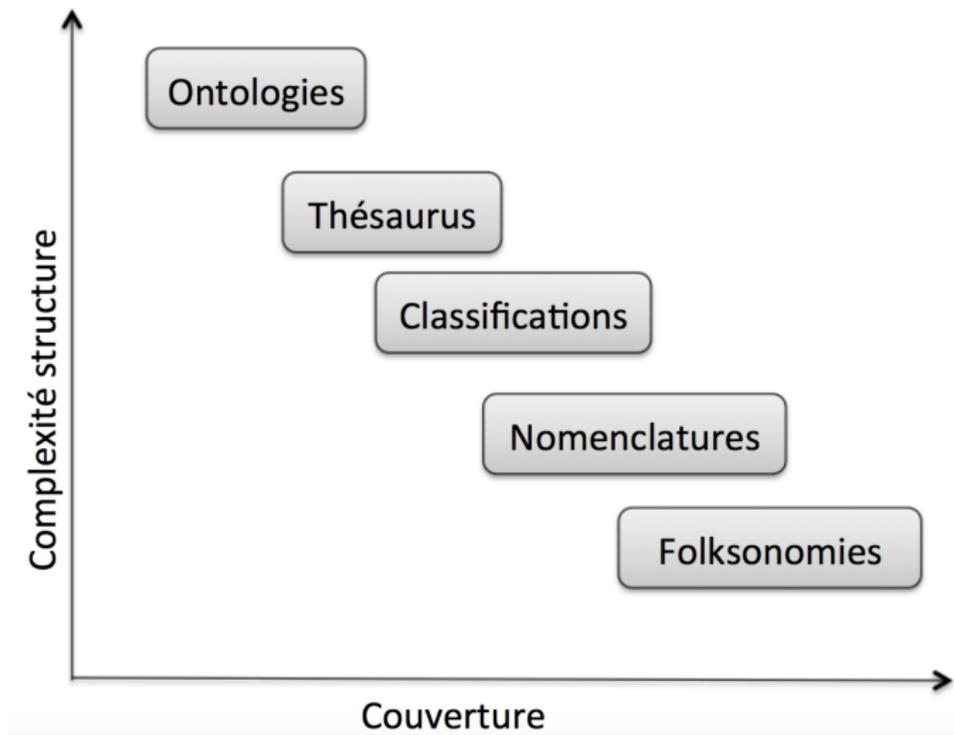
Notre approche pour la construction collaborative de connaissances factuelles et affectives

Conclusion

Description

- ▶ **Les SOC qui contiennent un vocabulaire contrôlé pour indexer les documents :**
 - ▶ les thésaurus,
 - ▶ les classifications et les
 - ▶ nomenclatures
- ▶ **Les ontologies :** étendent la structure formelle d'un **vocabulaire** en incluant plusieurs types de relations (synonymie, antonymie, etc.) entre les concepts.
- ▶ **Les folksonomies :** intègrent une nouvelle **dimension sociale** en donnant la possibilité à l'utilisateur d'indexer des documents, [Marlow, Naaman, Boyd & Davis, 2006]

Description



Plan

Introduction

Les systèmes d'organisation de connaissances

Le crowdsourcing pour la construction collaborative des connaissances

Notre approche pour la construction collaborative de connaissances factuelles et affectives

Conclusion

Le Crowdsourcing

- ▶ Le terme **crowdsourcing** est apparu en 2005 [Safire, 2009]
- ▶ Adopté rapidement par la communauté scientifique
- ▶ Les premières études présentent dans la littérature mentionnent **Amazon Mechanical Turk (AMT)**, avec **CrowdFlower** qui est apparu un peu plus tard.
- ▶ Plusieurs domaines de recherche ont expérimenté le crowdsourcing

Dans le domaine de la sémantique

- ▶ [Rumshisky et al., 2009](#) : ont expérimenté la construction d'une base de données contenant les sens flous d'un ensemble de verbes polysémiques.
 - ▶ Les annotateurs annotent les verbes dans leur contexte d'apparition.
 - ▶ **F-score de 0.93** en comparant les annotations effectuées par des lexicographes professionnels et celles effectuées par les annotateurs
- ▶ [Rumshisky, 2011](#) : conclut que le clustering de 350 lignes par groupe de sens similaires **n'est fiable que pour seulement environ 140 lignes**
- ▶ [Rumshisky et al., 2012](#) : expliquent que l'un des problème majeur est d'attirer des annotateurs avec un bon niveau de performance sur AMT pour effectuer des tâches linguistiques complexes.

Dans le domaine de la fouille d'opinions

- ▶ [Hsueh et al., 2009](#) : ont utilisé AMT pour la classification en polarité (positif, négatif, mixte ou neutre) des segments de texte extraits depuis des blogs politiques.
- ▶ Les auteurs mentionnent une valeur de **Kappa de 70.4%**.
- ▶ [\[Sayeed et al., 2011\]](#) : ont utilisé CrowdFlower pour une tâche d'annotation plus fine pour identifier le terme qui exprime une opinion par rapport à un terme cible présent dans la phrase.
- ▶ La meilleure valeur de **Kappa est de 0.65** obtenue lors de la suppression de 7 annotateurs douteux

Dans le domaine des entités nommées

- ▶ [Finin et al., 2010] : décrivent leur expérience d'annotation en entités nommées des messages Twitter
- ▶ La performance des annotateurs de AMT a été jugée **inférieure à celle des annotateurs experts** mais il est possible d'atteindre des performances équivalentes en combinant les annotations de plusieurs annotateurs.

Plan

Introduction

Les systèmes d'organisation de connaissances

Le crowdsourcing pour la construction collaborative des connaissances

Notre approche pour la construction collaborative de connaissances factuelles et affectives

Conclusion

Collecte du corpus textuel

- ▶ Parmi les objectifs du projet européen uComp :
 - ▶ La constitution d'une **base de connaissances factuelles et affectives** pour le domaine du **changement climatique**
- ▶ Collecte basée sur une liste fermée de mots clés

	Anglais	Français	Allemand
# Articles journalistiques	1 275 000	720 000	650 000
# messages médias sociaux	20 000 000	410 000	565 000

TABLE: Caractéristiques du corpus collecté [Weichselbraun et al., 2013]

Extraction des termes affectifs candidats

- ▶ Extraction automatique de termes affectifs pour chaque classe d'émotion décrite dans [Fraisse et Paroubek, 2014]
- ▶ Approche statistique : basée sur la mesure de l'information mutuelle introduite par [Fano, 1961], [Church and Hanks, 1990].

$$IM(m, e) = \log_2 \left(\frac{freq(m, e)}{freq(m) \cdot freq(e)} \right)$$

FIGURE: Information mutuelle entre un mot m et une émotion e

Validation des termes affectifs candidats

- ▶ Validation par le crowdsourcing
 - ▶ validation de la liste des termes affectifs français (9939 termes)

ignorance

Ce terme peut-il être utilisé pour exprimer une opinion, un sentiment ou une émotion ?

- Oui
- Non

Ce terme peut-il déclencher une opinion, un sentiment ou une émotion ?

- Oui
- Non

Ce terme a-t-il une connotation

- Positive
- Negative
- Neutre

Quelle est l'opinion, le sentiment ou l'émotion que ce terme peut exprimer ou déclencher ?

- AMOUR
- APPAISEMENT
- DERANGEMENT
- ENNUI
- PLAISIR
- DEPLAISIR
- TRISTESSE
- PEUR
- COLÈRE

FIGURE: Description de la tâche d'annotation

Expériences et Résultats

▶ **Expérience 1 :**

- ▶ Annotateur avec le **niveau Best** (meilleur niveau)
- ▶ Prix de la tâche fixé à **6 centimes**
- ▶ Cela nous a permis de valider uniquement 9,1% des termes

▶ **Expérience 2 :**

- ▶ Annotateur avec le **niveau Good** (le plus bas niveau)
- ▶ Validation de la liste de termes en une semaine

Jeu à but : Climate Challenge

- ▶ Jeu en cours d'expérimentation
- ▶ Dans un premier temps, la tâche de validation consiste à attribuer à un terme une valeur de polarité parmi les 5 valeurs suivantes : **très négatif, négatif, neutre, positif, très positif**
- ▶ Afin d'avoir un niveau de confiance assez élevé, le même terme est validé par plusieurs joueurs
- ▶ Chaque joueur dispose d'un **score** qui indique son niveau dans le jeu : ce score est calculé en fonction du nombre total de termes validés par le joueur ainsi que le nombre de bonnes réponses.

Jeu à but : Climate Challenge

Climate challenge

test your climate knowledge - predict the future state of our world



Play Progress Leaderboard About



Sentiment Assessment

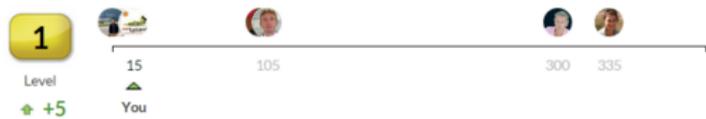
Does the following term have a **negative**, neutral or **positive** meaning?

energy conservation



Game Level

1
Level


15 You
105
300 335

Current Questions

OPEN QUESTIONS

- How much warmer or colder than average will sea surface temperature be in April 2015 in the eastern tropical Pacific Ocean (the Niño 3.4 region)?

Top Scores March

1		Christoph Meili Level: 3	1150
2		Diana Maynard Level: 3	950
8		Antonio Panderas Level: 1	100

Constitution collaborative d'une ontologie du domaine

- ▶ Projet uComp :
- ▶ [Wohlgenannt et al. \(2013\)](#) ont extrait une première liste de termes candidats à partir du corpus textuel collecté
- ▶ Utilisation de WordNet pour établir des relations (synonymie, antonymie, etc.) entre les concepts collectés
- ▶ Validation de connaissances construites par le crowdsourcing
 - ▶ Développement d'un plugin Protégé connecté à CrowdFlower

Validation de l'ontologie

The screenshot shows the Protégé software interface with the 'uComp Class Validation' plugin active. The left sidebar displays a class hierarchy starting with 'Thing' and including terms like 'carbon', 'climate', 'CO2', 'carbon dioxide', 'drought', 'earth', 'emission', 'energy', 'electricity', 'heat', 'heating', 'environment', 'evidence', 'flood', 'flooding', and 'fracking'. The 'energy' class is selected. The main window has tabs for 'Annotations', 'Usage', and 'uComp Class Validation'. The 'uComp Class Validation' tab is active, showing a form for validating the 'energy' concept. The form includes fields for 'Concept to be validated:' (filled with 'energy'), 'Validate relevance for domain:' (filled with 'Climate Change'), and 'Additional information for validators:' (filled with 'You can use external resources.'). There are radio buttons for 'Send to CrowdFlower' (selected) and 'Send to uComp-Quiz'. At the bottom, there is a 'Validate subtree' checkbox (checked) and buttons for 'CALCULATE COSTS', 'GO', 'Cancel subtree', and 'CAN'. A text box at the bottom of the form displays the results: 'This job contains **4 units**. You have set the number of judgements to **5** and the amount of money paid per task to **3 cents** per task. Expected total costs of this job are **\$0.60**.'

FIGURE: Interface du plugin Protégé pour la validation des termes de l'ontologie par les utilisateurs

Résultats

- ▶ Le résultat de cette approche a montré :
 - ▶ un score d'exactitude qui varie entre 89% et 99%,
 - ▶ que le temps passé est équivalent à celui d'un expert du domaine et
 - ▶ que le coût de production de l'ontologie est réduit de 75% à 81%.

Plan

Introduction

Les systèmes d'organisation de connaissances

Le crowdsourcing pour la construction collaborative des connaissances

Notre approche pour la construction collaborative de connaissances factuelles et affectives

Conclusion

Conclusions

- ▶ Nous avons proposé une approche basée sur la participation collaborative de différents types d'utilisateurs (simple utilisateur, experts du domaine, ingénieur de connaissance, etc.).
- ▶ Cette approche se base sur l'utilisation de jeu à but ainsi que le crowdsourcing.
- ▶ Nos deux expérimentations montrent que :
 - ▶ la qualité obtenue est équivalente à celle qui serait obtenue par des experts de la connaissance
 - ▶ les coûts et les délais sont réduits
- ▶ Cependant, le recours aux plateformes collaboratives comme AMT ou Crowdfunder pose d'important problèmes éthiques et juridiques (atelier sur l'éthique ETeRNAL à TALN 2015).

Merci pour votre attention !