Construction d'un large corpus libre de conversations écrites en ligne synchrones et asynchrones en français

Nicolas Hernandez et Soufian Salim

Université de Nantes – LINA CNRS UMR 6241

Rennes, JIR-CMC 2015

Contexte

Projet ODISAE 2014–2016 (`www.odisae.com`) soutenu par le FUI 17

Motivation applicative : Un analyseur sémantique de conversations écrites en ligne pour améliorer la gestion de la relation client (centre d'appel)

- recherche d'information inter-modalités textuelles
- mise à jour de FAQ (e.g. détecter l'absence d'une solution)
- aide à la génération de contenu (e.g. évaluer la complétude de messages réponses)
- génération d'alertes (e.g. énervement, attrition)

- des industriels avec des données à "enjeux réels"
- et des contraintes d'exploitation (e.g. diffusion) qui posent des problèmes pour un chercheur (i.e. reproduction et réutilisation)

Motivations scientifiques

Objet d'étude : les conversations (Médiées par les Réseaux) écrites multi-canaux

- mieux comprendre la structure et le fonctionnement de ces canaux d'expression ainsi que leurs interactions
- améliorer le traitement linguistique d'une modalité dégradée en exploitant la comparabilité inter-canaux (e.g. le langage "SMS" du chat)
- maîtriser l'alignement voire la "traduction" d'un contenu entre modalités textuelles (texte explicatif ou communication écrite en ligne)

Chercheur en TAL

- permettre des activités de Traitement Automatique des Langues (TAL) (e.g. construction de modèles statistiques)
- diffuser des corpus et des résultats d'analyse dans une perspective de reproduction et réutilisation (Nielsen, 2011)

Caractéristiques du corpus idéal

- conversations écrites en ligne
- multi-canaux i.e. synchrone (chat) et asynchrones (forum et courriel) et des textes explicatifs
- sur une même période
- représentatifs d'écrits récents
- même type de situation discursive à savoir l'assistance à la résolution de problèmes
- en français
- large
- voire en croissance perpétuelle
- libre

Dans la littérature, les corpus de CMR en français...

CoMeRe (Chanier et al., 2014) compiler, préparer et disséminer via OrtoLang

- 🙂exploiter des standards pour décrire données et méta-données
- 🙂réfléchir à l'exploitation (e.g. diffusion) d'un corpus en amont de sa construction
- 🙂adhésion au modèle de données TEI-CMR 🙁exploitation en TAL

Corpus Simuligne du projet LETEC (Reffay et al., 2014)

- multi-modalités (chat, courriel et forum) avec ressources pédagogiques
- autour de la situation d'apprentissage en ligne du français langue étrangère
- notre projet diffère en objectif, objet d'étude, situation discursive, quantité de données

Source du "corpus idéal"

Communauté du logiciel libre – Ubuntu-fr (francophone)

- forums de discussion
- listes de diffusion/discussion par courriel
- des canaux Internet Relay Channel (IRC)
- documentation

Démarche similaire pour la construction d'un corpus de chat en anglais (Uthus et al., 2013 ; Lowe et al., 2015)

Sommaire

Etat d'avancement dans la mise en place d'un cadre matériel pour soutenir notre recherche

- Collecte des données
- Considérations pour leur diffusion
- Définition des schémas et choix de format des méta-données, des données et des résultats d'analyse
- Traitements automatiques pour une tâche d'annotation manuelle en actes du dialogue

Licence d'exploitation

- accès public en consultation sur l'Internet
- documentation sous licence CC BY-SA v3.0
- absence de licence pour le forum, courriel et chat
  - droit d'auteur par défaut sur contenu des messages
  - usage délégué à l'éditeur Ubuntu-fr
  - Si utilisateur manifeste son refus de participer à ce corpus, nous devrons supprimer ses messages de notre copie
- accord pour le forum de ne pas faire indexer le contenu diffusé

Modes de collecte

doc.
- aucun archivage public et contenu mouvant
- aspiration quotidienne du site en ligne et versionnage
- travail avec la communauté pour systématiser archivage

courriel
- archive incrémentale distribuée publiquement

forum
- aucun archivage public et contenu peu mouvant
- aspiration incrémentale du site en ligne

chat
- le canal *fr* est non archivé
- procédure de journalisation depuis novembre 2014

Etat de la collecte

Près de 12 mois synchronisés

- 12 ans de forum et courriel disponible depuis leur création (à savoir 2004)
- chat et documentation depuis novembre 2014

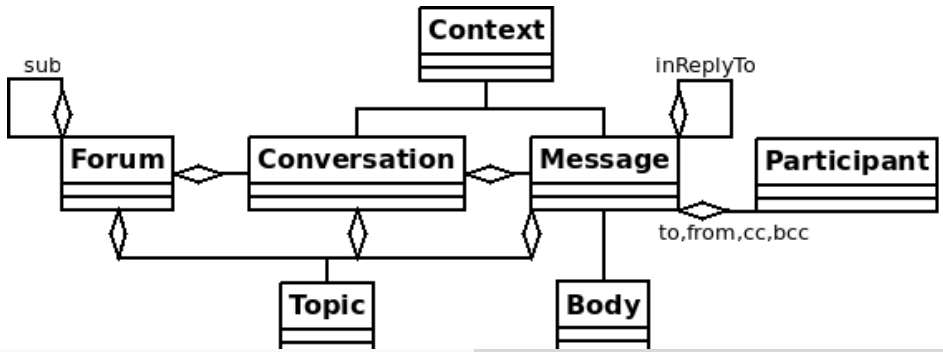
A titre indicatif, 6 mois de collecte à partir de nov. 2014

Modalité	**# conversations**	**# messages**	**# participants**
courriels	80	240	60
forums	12 000	90 000	7 000
chat		60 000	1 600
documentation	4 631 pages HTML et 4 301 618 tokens mots		

Modèle unifié pour décrire les méta-données...

... des conversations indépendamment du canal

- description d'une conversation, de ses messages et de ses participants
- se fonde sur généralisation des conversations observées et sur format des messages Internet (e.g. RFC6854)
- intègre des attributs communs et spécifiques pour catégoriser thématiquement une conversation, comptabiliser les vues d'un message, décrire le rôle d'un participant (e.g. ambassadeur, expert, client)
- un développement possible aux recommandations du projet CoMeRe



Informations présentes dans le modèle

- Forum/Room structure thématiquement les Conversations qui sont constitués de Messages. Structure récursive
- Conversations, Messages et Participants sont identifiées de manière unique
- Forum, Conversation, et Message peuvent avoir un sujet (e.g. "*Configuration Matérielle*", "*App ne détecte pas SD card*") et être catégorisés par des Topics (e.g. "*Hardware*", "*Phone*").
- Context d'un Message/Conversation : type de médium, statut privée, résolu, épinglé, compteur de likes/vu, degré d'importance degree,
- Message interconnecte Participants avec des relations from, to, cc, bcc
- Message sont constitués de Utterances associés chacun à une valeur OSE et acte du dialogue
- Body spécifie type mime et character encoding du contenu d'un message
- Conversation structures est exprimée via daytime et inReplyTo Message.

Format de sérialisation des annotations

Usage

- permettre à tous d'éditer et annoter la donnée dans sa mise en forme originale
- échanger et stocker

Requiert une couche d'abstraction vis-à-vis de la donnée

- adoption du principe de *"stand off" annotation* (aussi utile pour représenter annotations concurrentes et de type graphe)
- utilisation d'outils "conscients" e.g. Apache UIMA (Ferrucci et Lally, 2004), Webanno (Eckart de Castilho et al., 2014)

En pratique, sérialisation en XMI (*XML Metadata Interchange*), standard OMG pour échanges de données UML (*Unified Modeling Language*) en XML

On s'éloigne du format TEI mais pas de son modèle conceptuel

Représentation d'annotations en XMI

Ceci est un texte.
Et là une nouvelle phrase.

```
<?xml version="1.0" encoding="UTF-8"?><xmi:XMI xmlns:pos="http:///de/tudarmstadt/ukp/dkpro/core/api/lexmorph/type/pos
http:///uima/tcas.ecore" xmlns:xmi="http://www.omg.org/XMI" xmlns:cas="http:///uima/cas.ecore" xmlns:tweet="http:///d
dkpro/core/api/lexmorph/type/pos/tweet.ecore" xmlns:morph="http:///de/tudarmstadt/ukp/dkpro/core/api/lexmorph/type/mo
dependency="http:///de/tudarmstadt/ukp/dkpro/core/api/syntax/type/dependency.ecore" xmlns:type5="http:///de/tudarmsta
semantics/type.ecore" xmlns:type6="http:///de/tudarmstadt/ukp/dkpro/core/api/syntax/type.ecore" xmlns:type2="http:///
dkpro/core/api/metadata/type.ecore" xmlns:type3="http:///de/tudarmstadt/ukp/dkpro/core/api/ner/type.ecore" xmlns:type
tudarmstadt/ukp/dkpro/core/api/segmentation/type.ecore" xmlns:type="http:///de/tudarmstadt/ukp/dkpro/core/api/coref/t
constituent="http:///de/tudarmstadt/ukp/dkpro/core/api/syntax/type/constituent.ecore" xmlns:chunk="http:///de/tudarms
api/syntax/type/chunk.ecore" xmlns:custom="http:///webanno/custom.ecore" xmi:version="2.0"><cas:NULL xmi:id="0"/><cas
sofaNum="1" sofaID="_InitialView" mimeType="text" sofaString="Ceci est un texte.&#10;Et là une nouvelle phrase.&#10;"
DocumentMetaData xmi:id="1" sofa="12" begin="0" end="46" language="x-unspecified" documentTitle="texteSimple.txt" doc
documentUri="file:/media/hernandez-n/ext4/applications/webanno/datastore.150627/repository/project/14/document/723/so
collectionId="file:/media/hernandez-n/ext4/applications/webanno/datastore.150627/repository/project/14/document/723/s
documentBaseUri="file:/media/hernandez-n/ext4/applications/webanno/datastore.150627/repository/project/14/document/72
isLastSegment="false"/><type4:Sentence xmi:id="19" sofa="12" begin="0" end="18"/><type4:Sentence xmi:id="68" sofa="12
/><type4:Token xmi:id="23" sofa="12" begin="0" end="4"/><type4:Token xmi:id="32" sofa="12" begin="5" end="8"/><type4:
="12" begin="9" end="11"/><type4:Token xmi:id="50" sofa="12" begin="12" end="17"/><type4:Token xmi:id="59" sofa="12"
type4:Token xmi:id="72" sofa="12" begin="19" end="21"/><type4:Token xmi:id="81" sofa="12" begin="22" end="24"/><type4
sofa="12" begin="25" end="28"/><type4:Token xmi:id="99" sofa="12" begin="29" end="37"/><type4:Token xmi:id="108" sofa
44"/><type4:Token xmi:id="117" sofa="12" begin="44" end="45"/><type2:TagsetDescription xmi:id="126" sofa="12" begin="
tudarmstadt.ukp.dkpro.core.api.syntax.type.dependency.Dependency" name="Tiger"/><type2:TagsetDescription xmi:id="133"
end="0" layer="webanno.custom.FunctionalTextSegment" name="TenseQualifTagset"/><type2:TagsetDescription xmi:id="140"
end="0" layer="de.tudarmstadt.ukp.dkpro.core.api.ner.type.NamedEntity" name="NER_WebAnno"/><type2:TagsetDescription x
begin="0" end="0" layer="de.tudarmstadt.ukp.dkpro.core.api.lexmorph.type.pos.POS" name="STTS"/><cas:View sofa="12" me
41 50 59 72 81 90 99 108 117 126 133 140 147"/></xmi:XMI>
```

Préparation des données notamment pour une tâche d'annotation

Acte du dialogue

description d'un segment textuel en terme d'une fonction communicative (e.g. question, réponse, remerciement) relative à un contenu sémantique (e.g. activité, perception, obligation sociale) ; DIT++ (Bunt, 2009)

Etapes de pré-traitement

En fonction de la structure explicitée dans le support FluxBB, mbox, csv

1. Extraction des méta-données
   gestion des spécificités des supports e.g. encodage et MIME pour courriels ; destinataire non explicite dans les chats
2. Extraction du contenu textuel
3. Reconstitution des conversations
   champ *inReplyTo* et similarité de sujets pour les courriels et clustering lexical+information discursive pour le chat (Riou et al. 2015)
4. Segmentation en token mots et en énoncés
   règles sur traits surfaciques et mise en forme, spécialisé par canal, pour écrits des CMC et HTML

Anonymisation

Nous discutons plusieurs alternatives

(à replacer dans le contexte de l'existence d'un accès public)

- aucune anonymisation et suppression de messages à la demande
- anonymisation restreinte aux méta-données
- aucune diffusion de données, seulement des outils de collecte et de traitement des corpus (*à la Twitter*, mais c'est gênant pour le chat)

Connaître la forme des "logins" est intéressant...

Extrait d'une conversation de chat annoté en actes du dialogue

Quelques statistiques sur le corpus annoté

	# conversation	# message	# token	# dialog act
chat		2,320	17,448	1,989
forum	29	258	25,205	1,338
courriel	45	200	19,798	1,382

Distribution des dimensions selon les canaux

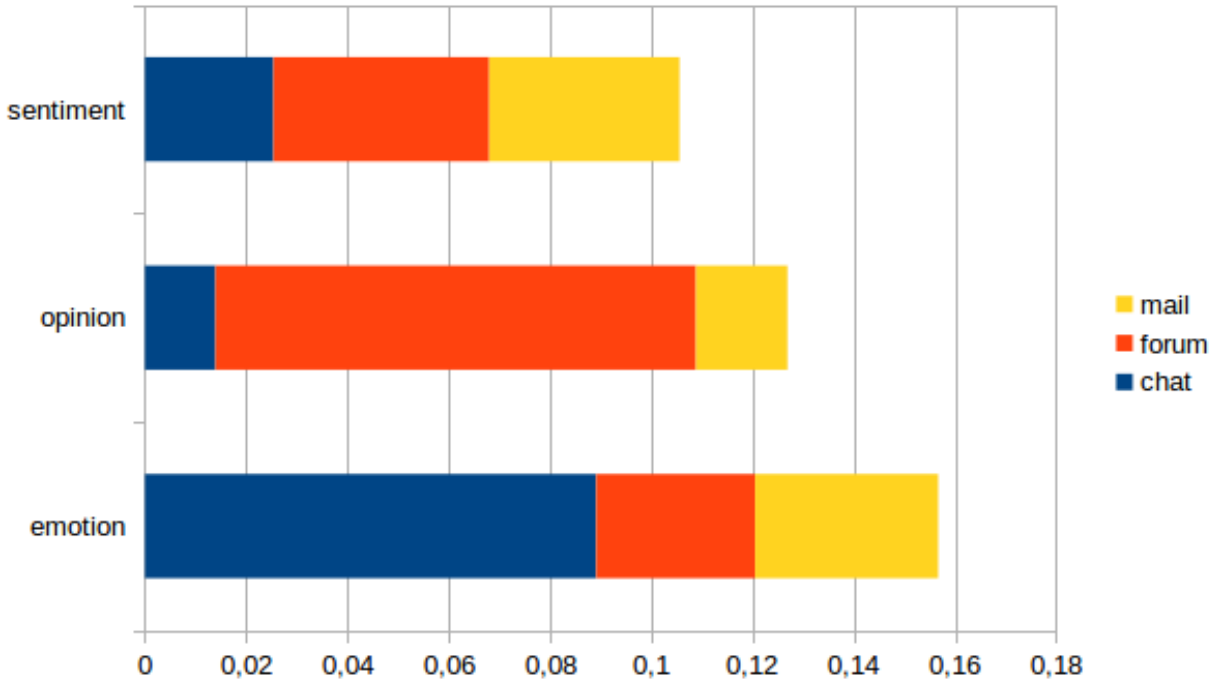
Dimension	chat	forum	mail
domainActivities	82,35	80,1	67
socialObligationManagement	9,25	12,85	30,35
discourseManagement	0,85	4,8	2
evaluation	3,95	1,65	0,15
psychologicalState	1,45	0,45	0,35
attentionPerceptionInterpretation	0,6	0,15	0,05
communicationManagement	1,35	0	0
contactManagement	0,9	0	0
timeManagement	0,2	0	0

Distribution des actes du dialogue selon les canaux

function	chat	forum	mail
inform	26,95	31,3	33,35
answer	17,65	20,05	11,2
requestForInformation	16,35	9,2	6,95
answerPositively	9,1	5,45	3,05
requestForAction	8,85	8,45	6,15
greetings	5,65	5,1	6,8
correct	4,75	3	1,3
answerNegatively	3,4	2,85	1,9
thanking	2,05	2,85	3,4
commit	1,95	1,05	1
requestForDirective	0,85	0,9	0,35
valediction	0,5	1,35	6,15
apologizing	0,45	0,65	0,6
anticipateThanking	0,4	1,95	2,95
finalSelfIntroduction	0	0,9	0,2

Distribution des opinions, sentiments et des émotions selon les canaux

Taxonomie fondée sur le projet ucomp [1] (Fraisse et al., 2014)



1. http://www.ucomp.eu

Conclusions

Proposition

- des conversations Médiées par les Réseaux
- un modèle unifié pour manipuler les conversations indépendamment des spécificités de chaque canal
- un large corpus libre de conversations écrites en ligne multi-canaux (synchrones et asynchrones) en français
- importance de déporter les annotations hors de la donnée (*"stand off"*)

Perspective

- Pour le corpus : son *packaging* et sa diffusion (Ortolang ?)
- Back to the scientific motivations

Merci pour votre attention
Des questions ?