Aide à l'automatisation de conception de systèmes de dialogue

Jean-Léon Bouraoui, Vincent Lemaire

Atelier - Fouille de Textes - Text Mine – *24/01/2017*

orange™

Sommaire

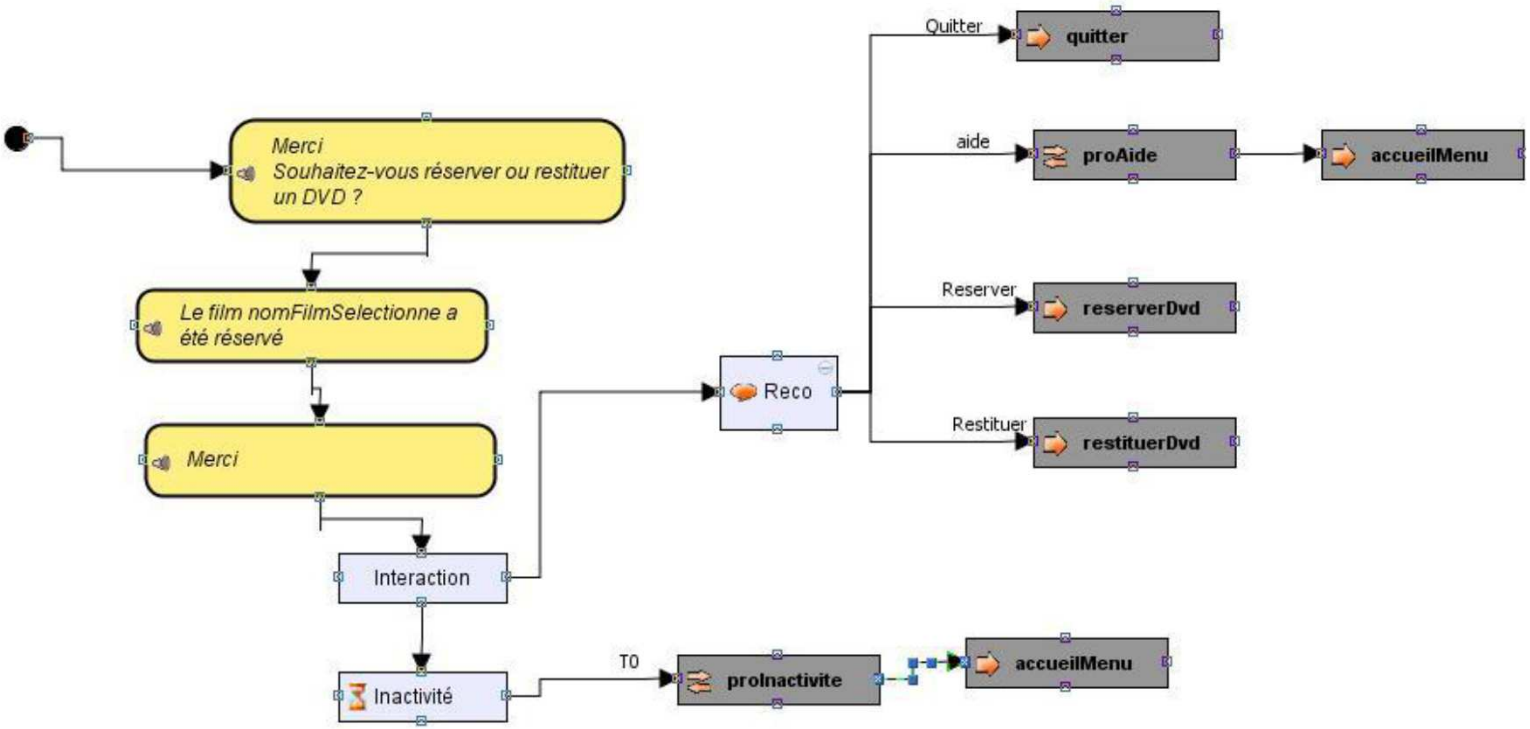
- Description de la problématique
- Etat de l'art
- Description de notre approche
- Corpus utilisés
- Illustrations sur un cas concret

Sommaire

- Description de la problématique
- Etat de l'art
- Description de notre approche
- Corpus utilisés
- Illustrations sur un cas concret

Problématique

Déduire une structure de dialogues à partir d'un ensemble de données textuelles (non supervisé)

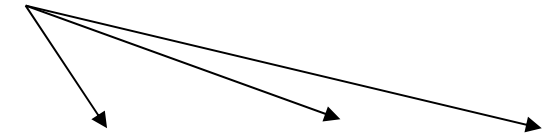


Modélisation semi-automatique des systèmes de dialogue

- Objectif: déduire une structure de dialogues à partir d'un ensemble de données textuelles (non supervisé)
  - A court terme : à partir de dialogues finalisés (homme – homme / homme – machine)
  - A moyen/long terme : à partir d'autres données (forums de discussions, FAQ, sites web, bases de données)

Modélisation semi-automatique des systèmes de dialogue

Exemple d'input :
- Dialogues de réservation de trains ; chaque dialogue se compose de plusieurs tours de parole, qui correspondent à différents thèmes (ici étiquetés « A », « B »). **NB** : « Agent » peut désigner un système de dialogue ou un humain répondant à l'usager



Dialogue 1 :
- User
- Agent
- User
- Agent
(…)
- User
- System

Dialogue 2 :
- User
- Agent
- User
- Agent
- User
- Agent
(…)

A

B

A

B

Exemples de tours de paroles sur ce thème :

User: Je voudrais un train vers 17h
Agent : Il y a le train de 17h15

User: Quels sont les trains entre 11h et 13h ?
Agent : Il y en a 2, un à 11h et l'autre à 13h

Exemples de tours de paroles sur ce thème :

User: Je voudrais partir de Paris
Agent : Départ de Paris enregistré

User: Est-ce qu'il vaut mieux partir de Paris ou du Mans
Agent : Je n'ai pas compris votre demande

Modélisation semi-automatique des systèmes de dialogue

Notation :

- $\mathbf{T_n}$: thème regroupant un à plusieurs tours de parole
- **User** : tour de parole d'un utilisateur humain
- **Agent** : tour de parole d'un système automatique (ou d'un autre humain)

Etape 1 : attribution d'un thème à chaque tour de parole

Dialogue 1 :

- User = $\mathbf{T_1}$
- Agent = $\mathbf{T_3}$
- User = $\mathbf{T_4}$
- Agent = $\mathbf{T_5}$

(…)

Dialogue 2 :

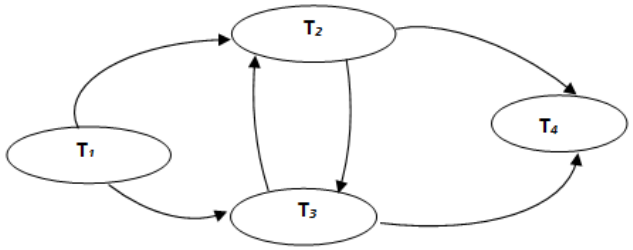
- User = $\mathbf{T_3}$
- Agent = $\mathbf{T_1}$
- User = $\mathbf{T_3}$
- Agent = $\mathbf{T_5}$
- User = $\mathbf{T_2}$
- Agent = $\mathbf{T_4}$

(…)

Dialogue *n* :

(…)

Etape 2 : calcul des transitions entre thèmes



Sommaire

- Description de la problématique
- **Etat de l'art**
- Description de notre approche
- Corpus utilisés
- Illustrations sur un cas concret

Etat de l'art

- Identification des thèmes puis des séquences de thèmes:
  - Clusters + HMM (Bangalore *et al.* (2008), Chotimongkol A. (2008))
  - LDA + HMM (Paul (2012), Zhai & Williams (2014)
- Deep Learning (Vinyals & Le Quoc (2015))
- Autres:
  - Actes de dialogues + CFG probabilisées (Alexandersson & Reithinger, (1997))
  - Heuristiques appliquées aux bases de données applicatives (D'Haro et al., 2009)
  - Intégration de techniques de Recherche d'Information dans le dialogue (Laroche (2015))
  - Clusters + heuristiques (Chalamalla *et al.* (2008), Negi *et al.* (2009))

Sommaire

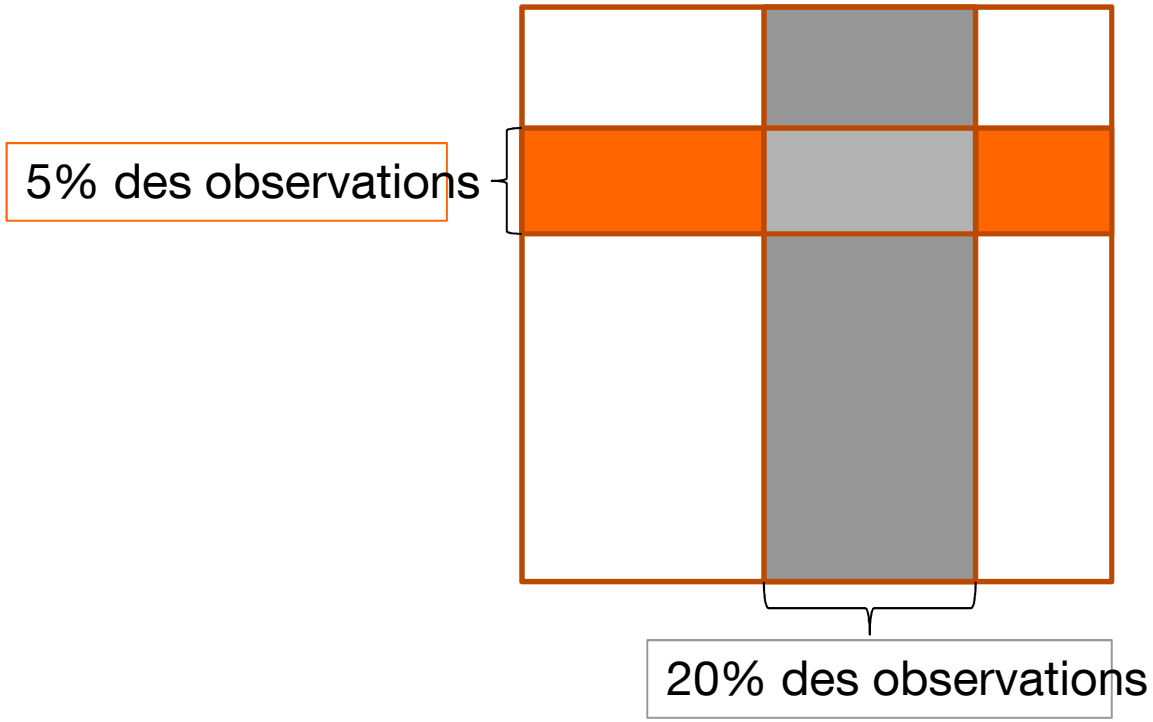
- Description de la problématique
- Etat de l'art
- **Description de notre approche**
- Corpus utilisés
- Illustrations sur un cas concret

Approche choisie

- Première version + Baseline pour de futures approches:
  - Clustering + calcul de fréquences

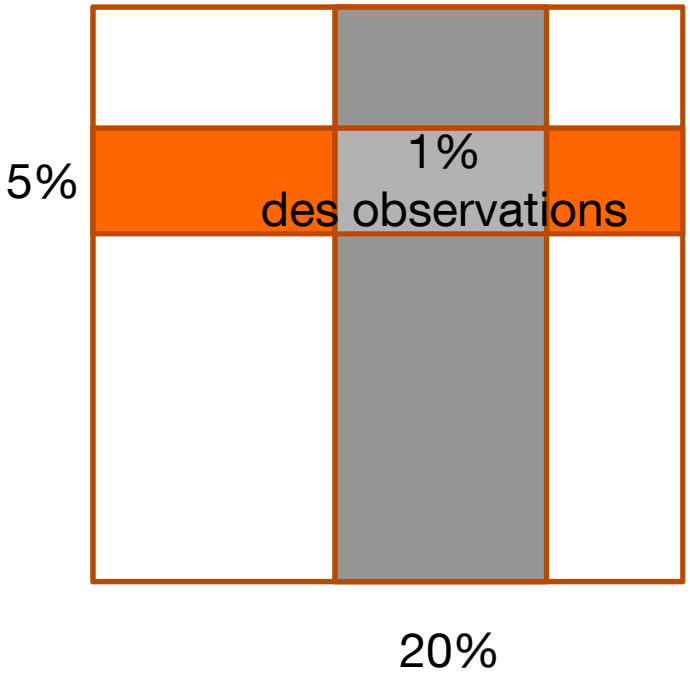
Qu’est-ce qu’un co-clustering ?

- Soit une co-partition :



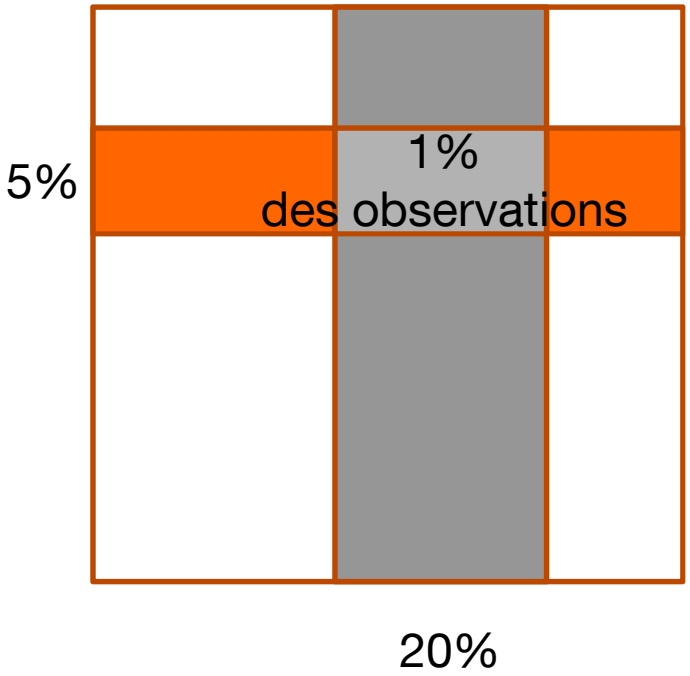
Qu'est-ce qu'un co-clustering ?

- Quelle information porte cette co-partie ?



Qu'est-ce qu'un co-clustering ?

- Quelle information porte cette co-partie ?
  - Aucune !
  - C'est ce qu'on attendrait sous l'hypothèse d'indépendance des variables (ligne, colonne)
    - Distribution des observations « au hasard » respectant les comptes des parties ligne et colonne (« marginales »)
      - 5/100 x 20/100 = 1/100 !



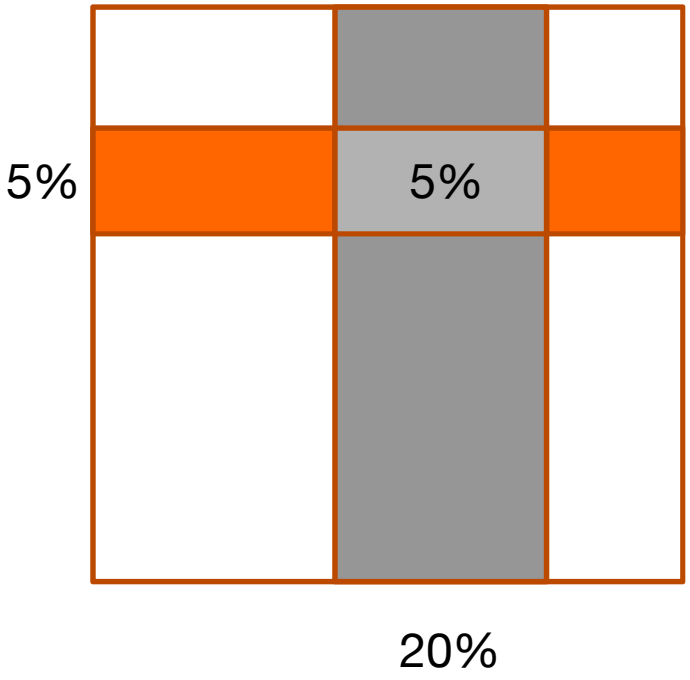
Qu’est-ce qu’un co-clustering ?

- Un co-clustering est une co-partition qui maximise le contraste entre la distribution des co-parties et la distribution espérée sous l’hypothèse d’indépendance (connaissant les marginales)

Qu'est-ce qu'un co-clustering ?

- Un co-clustering est une co-partition qui maximise le contraste entre la distribution des co-parties et la distribution espérée sous l'hypothèse d'indépendance (connaissant les marginales)

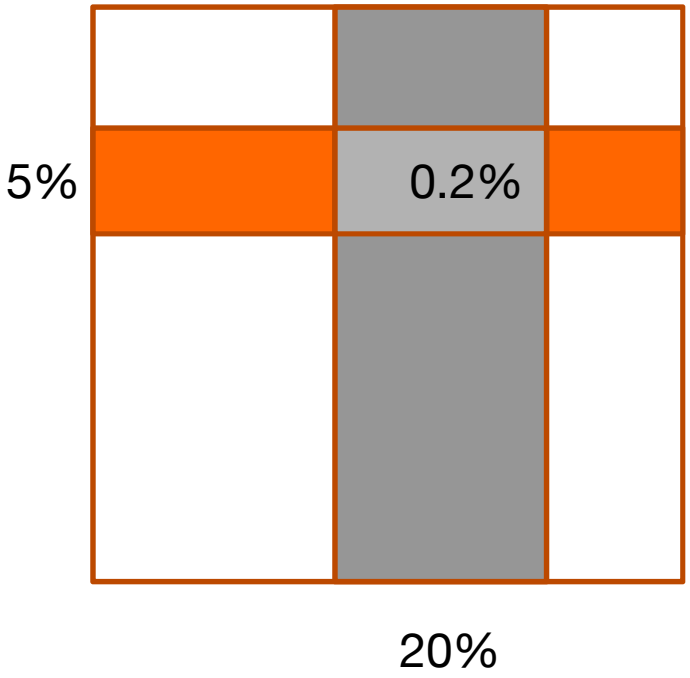
- Par exemple :



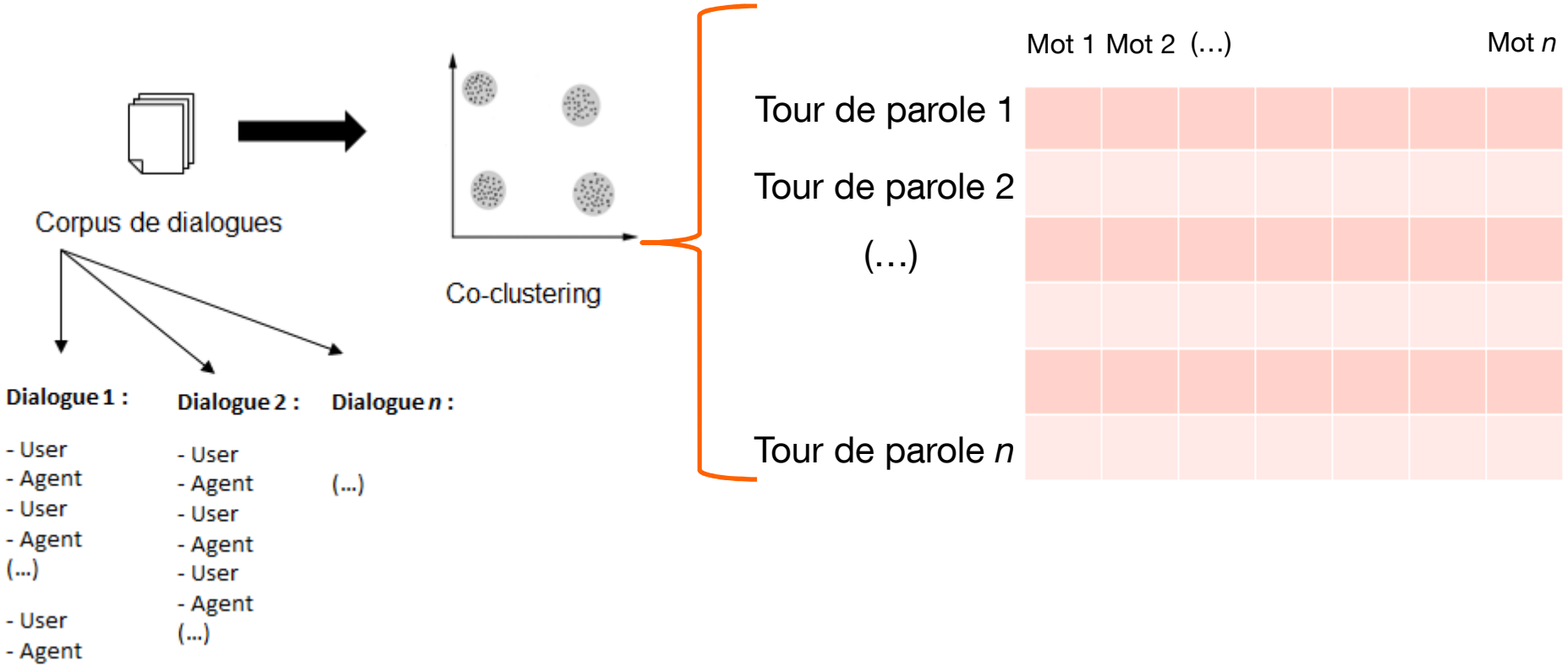
Qu'est-ce qu'un co-clustering ?

- Un co-clustering est une co-partition qui maximise le contraste entre la distribution des co-parties et la distribution espérée sous l'hypothèse d'indépendance (connaissant les marginales)

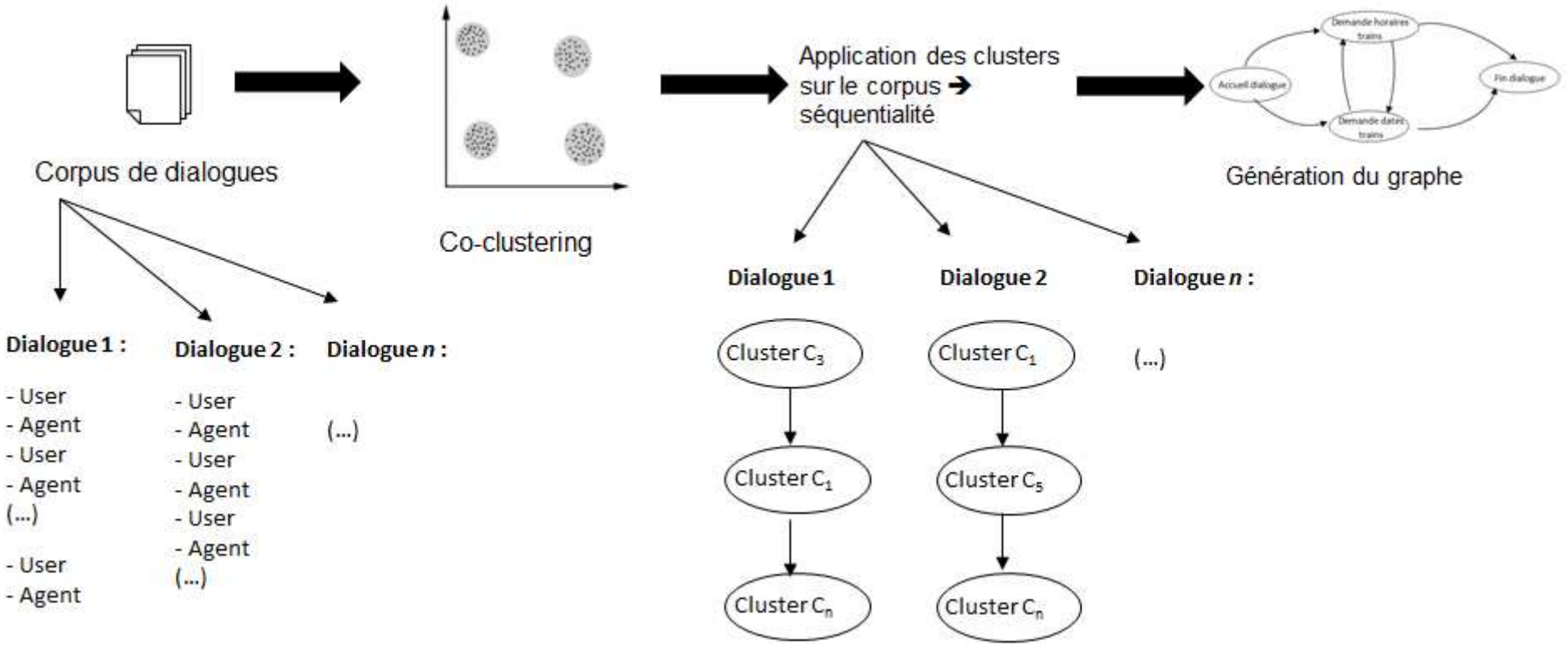
- Ou encore :



Chaîne de traitement



Chaîne de traitement



Sommaire

- Description de la problématique
- Etat de l'art
- Description de notre approche
- Corpus utilisés
- Illustrations sur un cas concret

Corpus utilisés

Origine	Air France	Orange	Orange
Domaine	Renseignements sur vols Air France	Renseignements/ assistance sur code PUK	Renseignements/ assistance sur services Roaming
Format	Transcription (XML TEI): • des tours de parole • de l'identité des locuteurs (client ou opérateur) • de phénomènes du dialogue oral spontané	Transcription (.txt) : • des tours de parole • de l'identité des locuteurs (client ou opérateur)	Transcription (.txt) : • des tours de parole • de l'identité des locuteurs (client ou opérateur)

Sommaire

- Description de la problématique
- Etat de l'art
- Description de notre approche
- Corpus utilisés
- **Illustrations sur un cas concret**

Corpus utilisé pour l'illustration

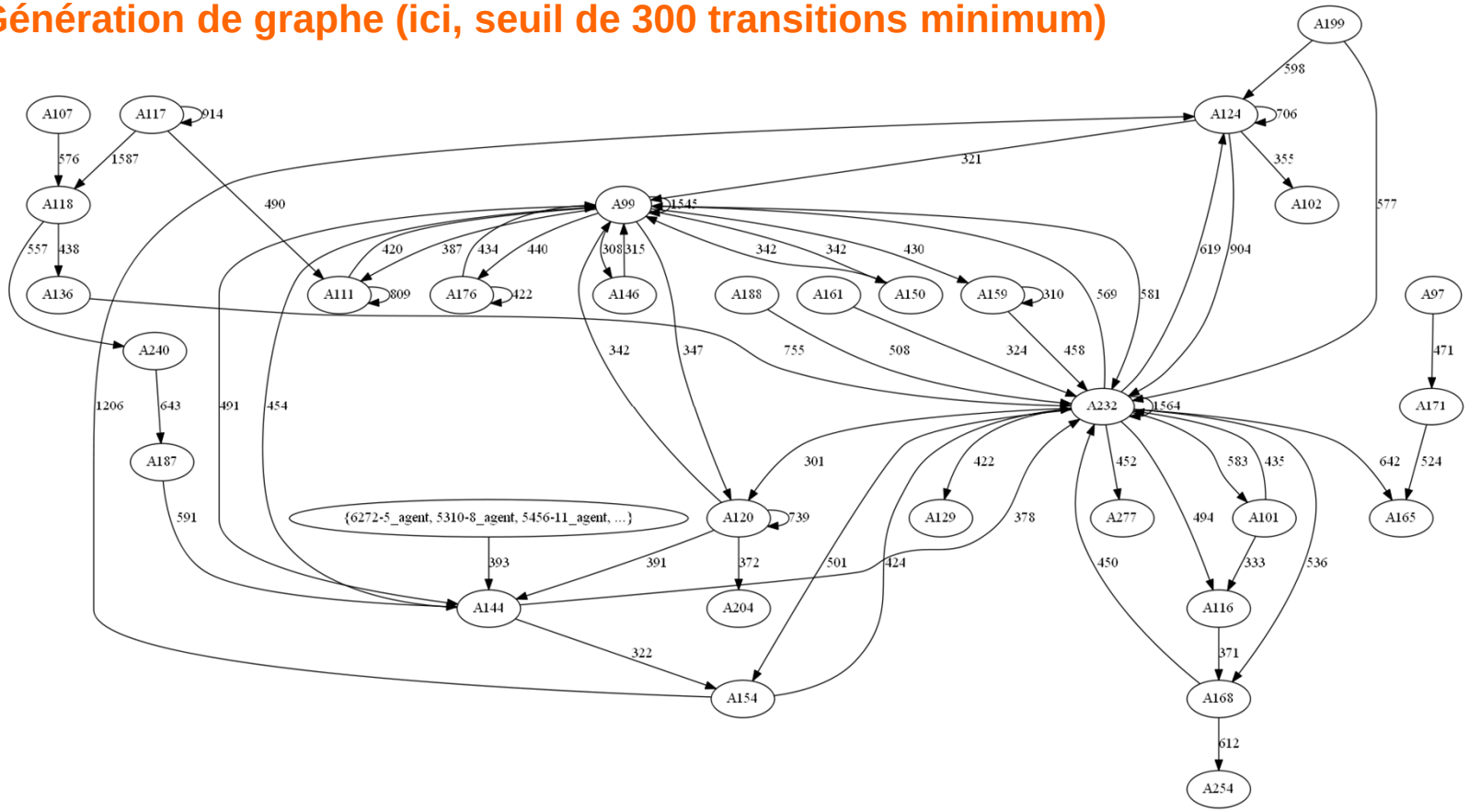
Orange
Renseignements/ assistance sur code PUK
Transcription (.txt) : • des tours de parole • de l'identité des locuteurs (client ou opérateur)

- 7407 dialogues différents
- Moyenne de 15 tours de parole / dialogue
- 37 318 mots différents
- ➔ Matrice ≈ (15*7407)* 37 318

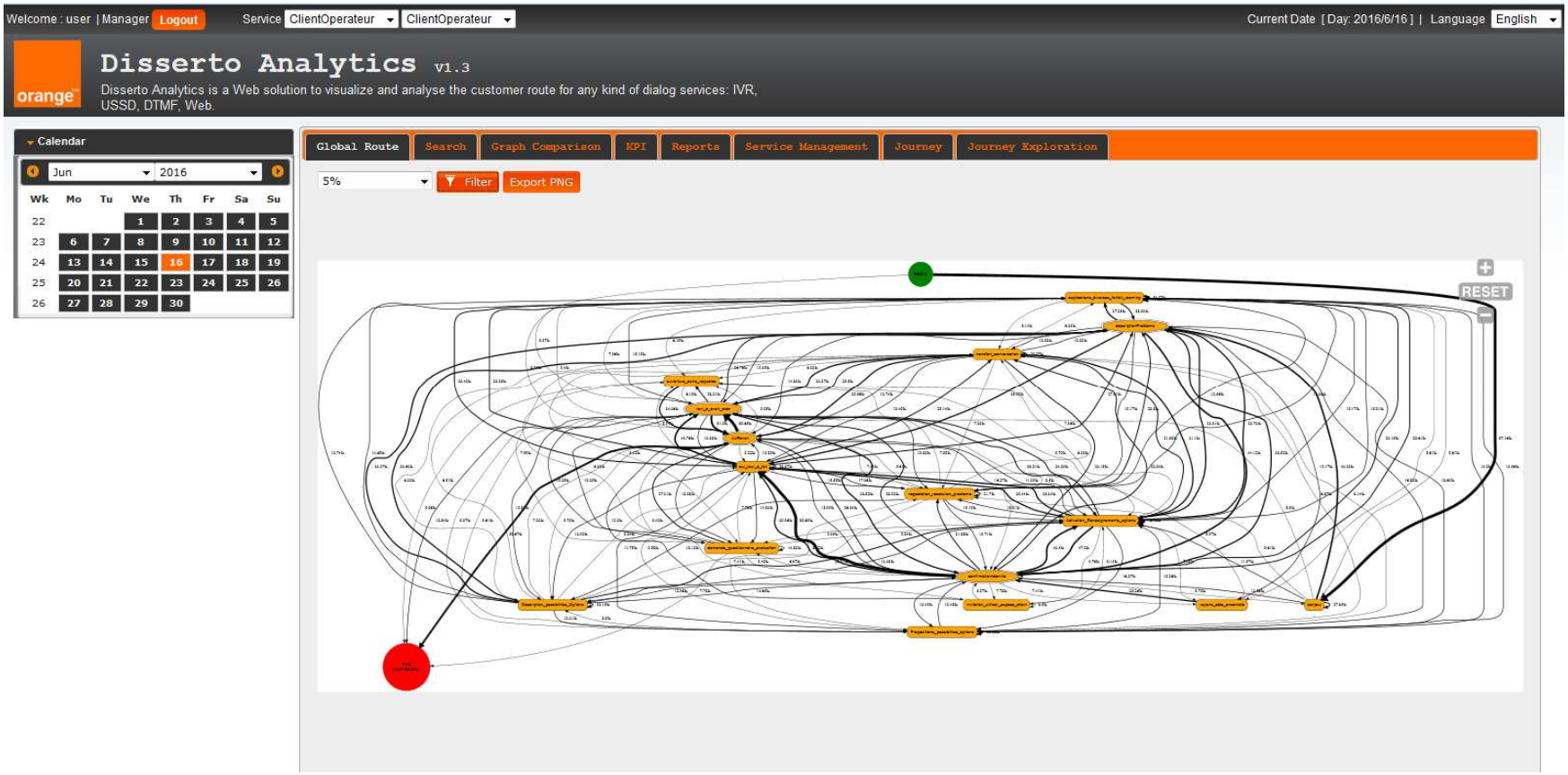
Résultats: premiers graphes

1. Pré-traitement du corpus
2. Co-clustering
3. Application des identifiants de cluster au corpus
4. Calcul des transitions

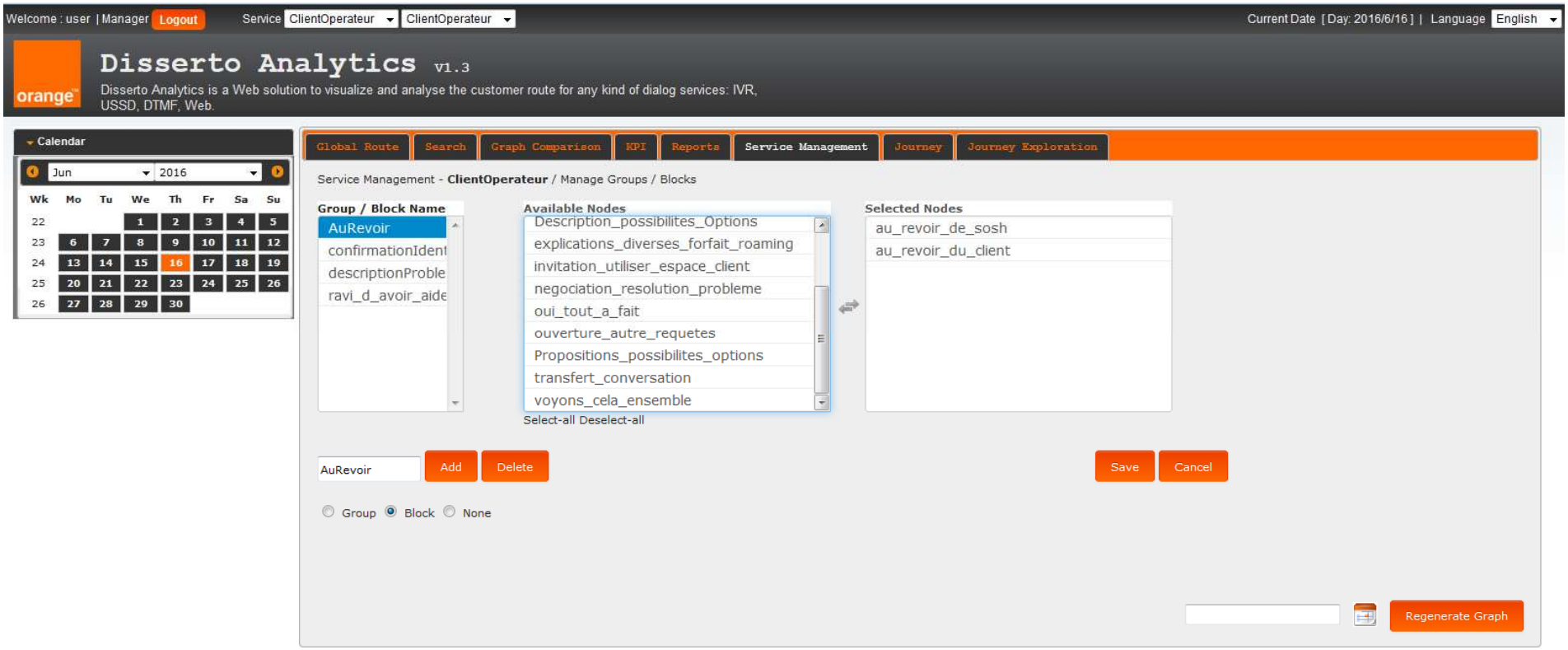
➔ **Génération de graphe (ici, seuil de 300 transitions minimum)**



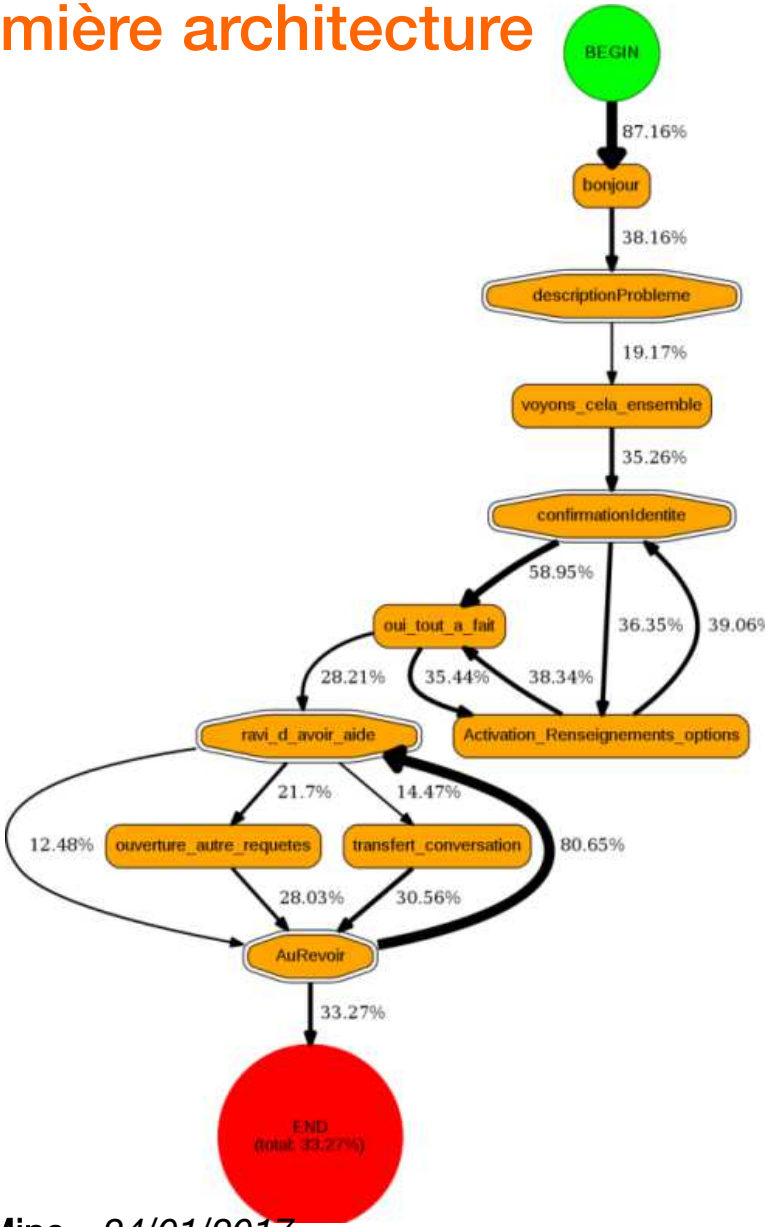
Résultats : visualisation avec Disserto Analytics



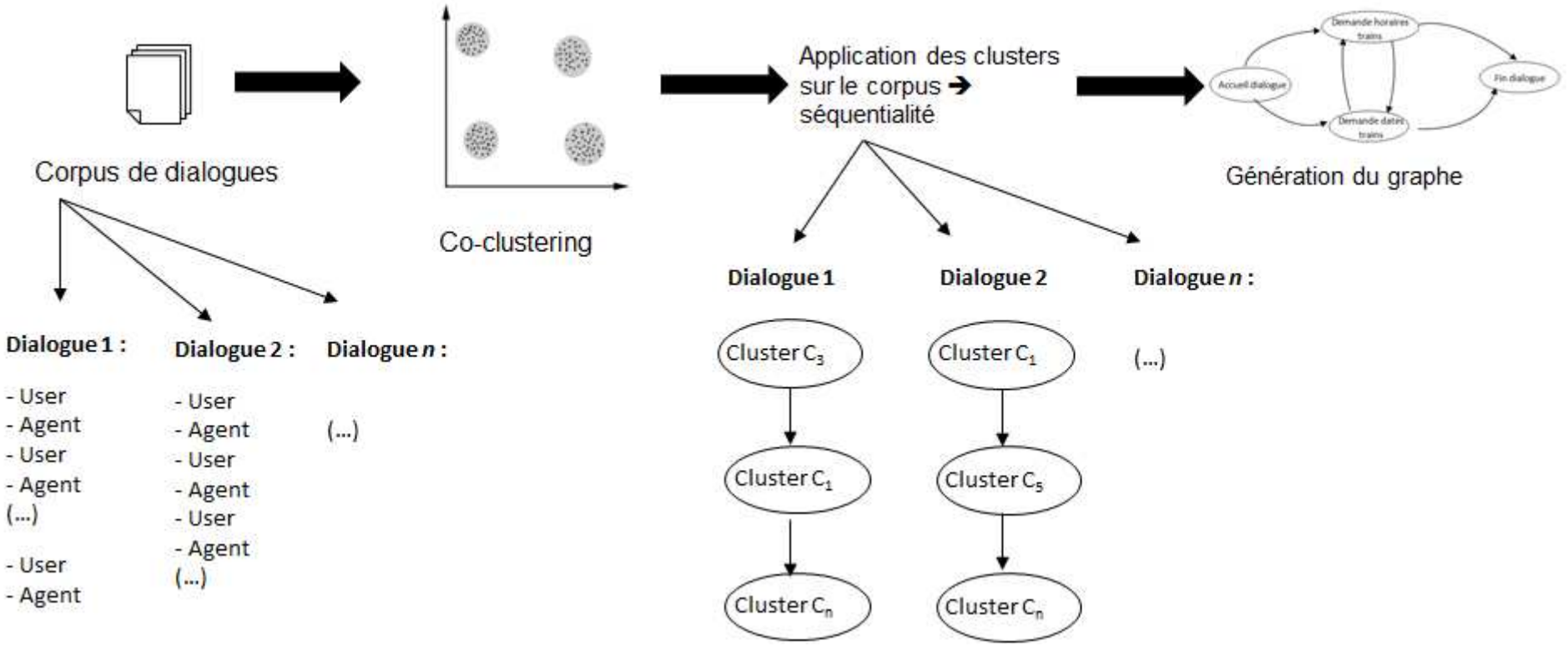
Résultats : visualisation avec Disserto Analytics



Résultats: première architecture



Conclusion


Corpus de dialogues
Co-clustering
Application des clusters sur le corpus → séquentialité
Génération du graphe
Dialogue 1 :
- User
- Agent
- User
- Agent
(...)
- User
- Agent
Dialogue 2 :
- User
- Agent
- User
- Agent
- User
- Agent
(...)
Dialogue n :
(...)
Dialogue 1
Cluster C3
Cluster C1
Cluster Cn
Dialogue 2
Cluster C1
Cluster C5
Cluster Cn
Dialogue n :
(...)

Perspectives

- Sélection des clusters les plus pertinents
- Thématisation des clusters
- Optimisations linguistiques sur le corpus
- Modélisation de la séquentialité ➔ HMM, CRF ?
- Cas de données réduites

Bibliographie

- Alexandersson and N. Reithinger, “Learning dialogue structures from a corpus”, *Eurospeech 1997*, pp. 8–1 5
- Bangalore S., Di Fabbrizio G., Stent A. (2008), «Learning the Structure of Task-driven Human-human Dialogs”, *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pp. 201–208
- Chalamalla A., Negi S., Joshi S., Subramaniam L. V. (2008), “Identification of Class Specific Discourse Patterns”, *CIKM '08*
- Chotimongkol A. (2008) *Learning the Structure of Task-Oriented Conversations from the Corpus of In-Domain Dialogs*, thèse de doctorat, Université Carnegie Melon
- D’Haro L. F., Cordoba R., Lucas J. M., Barra-Chicote R., San-Segundo R. (2009) “Speeding Up the Design of Dialogue Applications by Using Database Contents and Structure Information”, *SIGDIAL 2009*, pp.160–169
- Laroche R. (2015), "Content Finder AssistanT", *18th International Conference on Intelligence in Next Generation Networks,* pp. 231-238
- Negi S., Joshi S., Chalamallay A., Subramaniam L. V. (2009), “Automatically Extracting Dialog Models from Conversation Transcripts”, *2009 Ninth IEEE International Conference on Data Mining*
- Paul M., “Mixed membership Markov models for unsupervised conversation modeling”, *EMNLP-CoNLL '12*, pp. 94-104
- Vinyals O., Le Quoc. V (2015) “A neural conversational model”, *International Conference on Machine Learning*, Lille, France, 2015
- Zhai K., Williams J. (2014) “Discovering Latent Structure in Task-Oriented Dialogues”, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pp. 36–46,Baltimore, Maryland, USA, June 23-25 2014.

Merci!