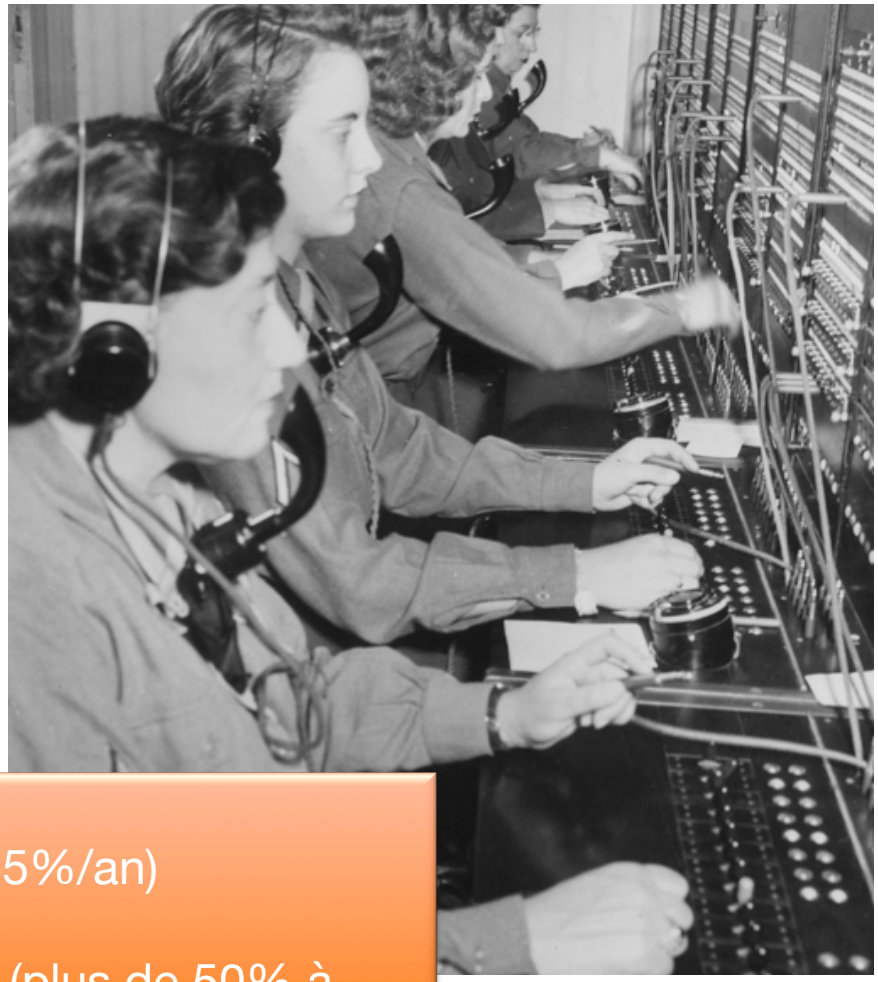
L’anonymisation, pierre d’achoppement pour le traitement automatique des courriels

Hugues de Mazancourt, Eptica
Alain Couillault, Aproged / L3I – Université de La Rochelle
Gaëlle Recourcé, Kwaga

Journée d’Etudes de l’ATALA, 22 Novembre 2014

Domaine : le support client

- La Gestion de la Relation Client représente un enjeu majeur d'image et de service
- Coût humain élevé
- Les plateformes logicielles actuelles se focalisent sur l'aspect « mise en relation » (avec un agent)
  - → peu d'utilisation du contenu des interactions



Contexte

- Augmentation continue des interactions (+5%/an)
- Multiplication des canaux
- Prééminence des interactions numériques (plus de 50% à horizon 2015)

De l’intérêt d’analyser le contenu

« In looking at our customer support data, he saw that free trial users who emailed us for support were nearly nine times more likely to convert to paying customers at the end of their trials than those who never reached out. »

Alex Turnbull, CEO & Founder of Groove

Le projet ODISAE

- Projet de R&D collaborative dans le cadre du FUI-17
  - Soutenu par les pôles Cap Digital et Images & Réseaux
  - Financé par BPIFrance et la Région Ile de France
- 9 Partenaires
  - Eptica
  - Kwaga
  - Jamespot
  - Cantoche
  - LINA
  - Aproged
  - TokyWoky
  - Centre Départemental du Tourisme de l'Aube
  - INSEE Contact
- Durée : 2 ans
- Budget total : 2,3 M€
- Financement : 993 k€

Eptica®

KWAGA

Jamespot
INTELLIGENCE . DAILY

CANTOCHE

lina

Aproged

TokyWoky

Aube en Champagne

Insee
Institut national de la statistique et des études économiques
Mesurer pour comprendre

cap·digital
Paris Region

ir
images& réseaux

bpifrance

îledeFrance

Objectifs du projet

- Etudier les conversations qui s'établissent entre le client et le(s) agent(s)
  - Pour les centres très performants, un problème est résolu dans 80% des cas
  - Mais ce chiffre est rarement atteint
  - Quoi qu'il en soit, il reste une proportion non-négligeable de cas dans lesquels une véritable conversation s'établit
- Le projet ODISAE porte donc sur l'analyse automatique des conversations
- Pour améliorer les systèmes de relation client
  - Eptica en premier lieu
- L'analyseur de conversations sera diffusé en open-source par le LINA

Ce que l'on va analyser

- La façon de s'adresser aux clients (resp. agents)
  - Croisée avec les thématiques
  - *« quand on parle de tondeuse B-32, ça tourne une fois sur deux au vinaigre »*
    - Problème sur le produit, sur la documentation,
    - Problème de compétence des agents
    - …
- Les intentions des clients
  - Intentions positives (achat) mais aussi négatives (procès, mauvaise publicité…)
  - À pondérer en fonction du profil du client (râleur ou non, etc.)
- La complétude des réponses
  - En cas de questions complexes, a-t-on répondu à tous les points ?

- Il faut pour cela un vaste corpus d'interactions avec des locuteurs identifiables

Pour faire un analyseur de conversations…

- Problème #1 : le corpus n'appartient pas aux partenaires technologiques
  - INSEE Contact a un historique d'échanges avec ses clients
  - Eptica a sollicité certains autres clients pour obtenir du corpus
- Problème #2 : il est interdit de diffuser un corpus sans l'autorisation explicite des auteurs
  - Droit d'auteur (ils sont plusieurs dizaines de milliers)
  - Droit à la correspondance privée, dont la violation est punie par 45 000€ d'amende et 1 an de prison

- Il n'y a de fait pas de corpus important de mails disponible pour une étude scientifique

Les mauvaises solutions

- Le corpus ENRON
  - Anglais
  - Vieux de plus de 10 ans
    - → outils, usages et modes d'expression datés
  - Contexte d'échange spécifique
- Les corpus de courriels « faits main »
  - Généralement dans le cadre de travaux de recherche, par collecte de courriels auprès des collègues/étudiants et anonymisés/pseudonymisés
    - Filtrage *a priori* des locuteurs et des contenus
    - Faible volume
- Les pseudo-corpus
  - Réalisés dans le cadre de scénarios qui simulent une situation réelle
    - Les acteurs ne sont pas les vrais acteurs
    - Les scénarios ne sont pas les situations réelles
- Les forums
  - Contexte d'échange très différent (on écrit pour être publié)
  - Pas toujours aussi libres qu'on le croit
    - *« il vous est interdit de procéder à une extraction qualitativement ou quantitativement substantielle des bases de données mises en ligne sur le site Doctissimo »* (Charte d'utilisation de doctissimo.fr)

Les préconisations européennes

- Emises par le G29 (« union des CNIL européennes ») sur la valorisation de données nominatives
  - Plus précisément sur le processus d'anonymisation préalable
- Les critères : il doit être impossible de
  - Isoler un individu
  - Corréler des ensembles de données distinctes concernant le même individu
  - Déduire de l'information sur un individu
- Si tous ces critères sont remplis, le corpus est considéré comme anonyme a priori
  - Sinon, il faut une « analyse détaillée des risques… »
- Inapplicable à notre contexte
  - Et à bien d'autres…

Les niveaux d'anonymisation dans ODISAE

- Niveau 1: effacement
  - On supprime toutes les données personnelles
  - → inutilisable pour une analyse de conversations
- Niveau 2: pseudonymisation
  - Suppression de l'ensemble des données non-utilisables pour le traitement linguistique (headers des mails…)
  - Remplacement données sensibles par des données similaires
    - Nom, prénom, adresse e-mail, mais aussi numéros de commande, etc.
  - → Bloque toute identification directe dans le monde réel
    - Personnes comme entreprises (revente de fichiers, image, …)
- Niveau 3: anonymisation vraie
  - Ré-écriture de certaines parties du texte (« obfsucation »), insertion/suppression de passages non-liés à la tâche ciblée
    - A priori étape manuelle
    - Donc très lourde
  - → C'est le seul niveau permettant la diffusion externe

Le fonctionnement dans ODISAE

- Les partenaires sont liés par un accord de confidentialité strict
- Les données fournies par les partenaires et clients sont préalablement nettoyées et anonymisées « niveau 1 »
  - Les données sont traitées chez le fournisseur avant d'en sortir
- Elles peuvent ensuite être transmises au consortium
  - Le niveau d'anonymisation est considéré comme acceptable par les fournisseurs de corpus
    - En particulier l'INSEE

- Les linguistes peuvent donc travailler au sein du consortium… mais pas publier les données

Pseudonymisation

```
Bonjour

Sauf erreure de ma part, je n'ai
pas reçu la facture de mon abris
de jardin Polka
Merci de me l'envoyer au plus vite
s'il vous plait, même par mail
j'ai rendez vous avec mon expert
comptablme jeudi

Dans cette attente

Cordialement

Pierre Monquignon
Les Jardin de Pays

> Message du 22/08/12 09:57
> De : "Service Clients Bricolo"
> A :  "monquignon"
```

```
Bonjour

Sauf erreure de ma part, je n'ai
pas reçu la facture de mon abris
de jardin Kroug
Merci de me l'envoyer au plus vite
s'il vous plait, même par mail
j'ai rendez vous avec mon expert
comptablme jeudi

Dans cette attente

Cordialement

Simon Cussonet
Zorglub Corp

> Message du 22/08/12 09:57
> De : "Service Clients Brol"
> A :    »cussonet"
```

Conclusion

- Nous avons défini une procédure permettant un travail collaboratif sur des données à caractère nominatif
  - Traitement automatique
  - Accord contractuel
- Permet le travail scientifique en consortium
  - Partage des corpus
  - Production de modèles qui vont travailler sur des « vraies » données
- Mais ne permet pas la diffusion massive de corpus d'interactions
  - Nécessiterait un travail humain ou une tâche automatique dédiée
  - Qui la rendrait peut-être conforme aux préconisations du G 29