Utilisation de contextes pour la correction automatique ou semi-automatique de réclamations clients

Philippe Suignard[1] Sofiane Kerroua[2]

(1) Electricité de France R&D, 1 avenue du Général de Gaulle, 92141 Clamart
(2) A.I.D., 4 rue Henri Le Sidaner, 78000 Versailles
philippe.suignard@edf.fr, skerroua@aid.fr

RÉSUMÉ

Cet article présente deux méthodes permettant de corriger des réclamations contenant des erreurs rédactionnelles, en s'appuyant sur le graphe des voisins orthographiques et contextuels. Ce graphe est constitué des formes ou mots trouvés dans un corpus d'apprentissage. Un lien entre deux formes traduit le fait que les deux formes se « ressemblent » et partagent des contextes similaires. La première méthode est semi-automatique et consiste à produire un dictionnaire de substitution à partir de ce graphe. La seconde méthode, plus ambitieuse, est entièrement automatisée. Elle s'appuie sur les contextes pour déterminer à quel mot correspond telle forme abrégée ou erronée. Les résultats ainsi obtenus permettent d'améliorer le processus déjà existant de constitution d'un dictionnaire de substitution mis en place au sein d'EDF.

ABSTRACT

Using contexts for automatic or semi-automatic correction of customer complaints

This article presents two methods allowing correcting complaints containing spelling errors, by using the spelling and contextual neighbors' graph. This graph is made of forms or words found in a learning corpus. A link between two forms conveys the fact that the two forms "look alike" and share similar contexts. The first method is semi-automatic and consists in producing a substitutional dictionary from this graph. The second method, more ambitious, is fully automatic. It is based on contexts to determine to which word corresponds such abbreviated or erroneous form. The results thus obtained allow us to improve the existing process regarding the creation of a substitutional dictionary at EDF.

MOTS-CLÉS : Correction automatique, analyse distributionnelle, graphe, contexte
KEYWORDS : Spelling correction, distributional analysis, graph, context

1 Introduction

Au sein des entreprises, un suivi et une analyse rigoureuse des réclamations, de leurs causes, et de leurs évolutions est une plus-value dans la connaissance du client. Cette problématique est rencontrée chez EDF qui analyse, rigoureusement, les réclamations, orales ou écrites, par le biais d'une chaîne de traitement. Celle-ci, prend sa source au sein des « *Centres de Relation Clientèle* » où sont recueillies, suivies et traitées toutes les demandes ou réclamations par les conseillers clientèles. Ceux-là ont la tâche d'accueillir le client, directement de vive voix ou par téléphone, indirectement par mail ou par

courrier, de déterminer les causes de leur requête, d'en apporter une solution, ou à défaut d'en avoir une, de contacter tous les services potentiellement capables de le faire, tout en prenant soin de maintenir le client satisfait des services offerts par leur fournisseur d'énergie et en lui proposant des offres commerciales. Ainsi, en plus de ces tâches, le conseiller doit saisir et décrire la réclamation du client. Dans ce contexte, les réclamations saisies par le conseiller sont sujettes à des erreurs rédactionnelles qu'il convient de corriger et de normaliser pour améliorer la qualité des traitements ultérieurs.

La suite de cet article décrit plus précisément les réclamations et leur analyse au sein d'EDF, ceci permettant de présenter le corpus d'apprentissage utilisé dans les parties suivantes. La partie 3 présente un état de l'art de la correction automatique de texte. La partie 4 présente les deux méthodes proposées pour la correction automatique. Toutes deux ayant pour pré-requis commun, la construction automatique du réseau des voisins orthographiques et contextuels. La partie 5 présente quelques résultats.

2 Les réclamations au sein d'EDF et le corpus d'apprentissage

En traitant les appels, les conseillers saisissent les réclamations des clients en y ajoutant des informations complémentaires (si le client avait déjà appelé, état de sa satisfaction, réponse apportée, etc.). Rédigée lors de l'appel, dans un cadre et dans un temps imparti et sans relecture *a posteriori*, la qualité de la réclamation est tributaire du conseiller qui la rédige. Ainsi, certaines réclamations, mal orthographiées et abrégées à outrance sont difficilement compréhensibles. De plus, le vocabulaire utilisé, abondamment abrégé, y est très spécialisé.

En France métropolitaine, on dénombre ainsi environ 200 000 réclamations par mois, exploitées, traitées et analysées par la Direction Commerce d'EDF, permettant ainsi de suivre l'évolution des demandes des clients.

Dans le but d'améliorer et de faciliter leur analyse, ces réclamations lors de leur traitement subissent une phase de normalisation qui consiste à remplacer des formes abrégées ou considérées comme erronées par des formes considérées comme étant canoniques. Formes canoniques, abrégées et erronées sont réunies dans un dictionnaire dit de substitution qui est utilisé lors de la normalisation.

Formes canoniques	Formes à corriger		
agence en ligne	ael	a.e.l	a-e-l
agent	agt		
alimentation	alim	alimention	

TABLE 1 - Extrait du dictionnaire de substitution

Ce dictionnaire de substitution est construit manuellement et enrichi au fil du temps par un expert métier ayant une bonne connaissance de la typologie orthographique des réclamations. Comme le montre la Table 1 , il s'agit d'un document texte tabulé où chaque ligne commence par la forme canonique et est suivie par une ou plusieurs formes abrégées ou considérées comme erronées. Cependant, ce dictionnaire, du fait de sa construction manuelle, ne peut pas être complet, la masse très importante des

commentaires ne permettant pas d'estimer, même pour un expert, la majorité des fautes ou des abréviations.

Nos travaux s'appuient sur un corpus d'apprentissage, composé de réclamations contenant un total de plus de 7 millions de mots. Les réclamations sont récupérées sans prétraitement, il s'agit donc des textes directement saisis par les conseillers.

Pour les tests qui suivent nous avons constitué un corpus appelé « corpus 100k » comprenant 100 000 réclamations.

3 Etat de l'art de la correction de texte

La correction de texte est un sujet qui a fait l'objet de nombreux brevets et travaux et qui continue à progresser du fait de l'évolution des moyens de production des textes (textes scannés, saisis avec des claviers d'ordinateur, puis des claviers de téléphones, etc.) et les contraintes associées (160 caractères pour les SMS ou 140 pour les tweets).

Beaucoup d'auteurs se sont penchés sur la problématique de la correction de texte. La plupart d'entre eux comme (Bouraoui *et al.*, 2009), commence par définir quelles sont ces erreurs et en établit une typologie, typologie que nous partageons largement. Notre corpus comprend :

- des inversions, ajouts ou suppressions de caractères (« cleint », « clint », « cllient » pour « client », suppression des « ç » comme dans « recu » ou des « è » comme dans « cheque ») ;
- des abréviations, formes raccourcies ou non terminées (« logt » pour « logement », « inter » pour « intervention », « pq » pour « pourquoi ») ;
- des sigles et acronymes (« mes » pour « mise en service ») ;
- des textes coupés en deux (« suite a ppel client ») ;
- des textes accolés ou agglutinés (« lavoir » pour « l'avoir », « le clienta » pour « le client a ») ;
- des textes coupés et accolés (« clienta ppel » pour « client appelle », « le client ma pel car… » pour « le client m'appelle car… ») ;
- des écritures phonétiques de type SMS (« ét » pour « été », « koi » pour « quoi », « 1client » pour « un client »)
- et bien sûr des fautes d'accord, de grammaire…

Ensuite, quelle méthode utiliser ? Marion Baranès (Baranès, 2012) en dresse un très large panorama : méthodes basées sur des dictionnaires, sur des règles de grammaires, méthodes utilisant les mots cooccurrents, méthodes utilisant différentes mesures de proximité (lexicale, clavier, phonétique, notamment pour corriger les SMS), classification, utilisation des n-grammes, etc. D'autres méthodes sont des combinaisons de toutes ces méthodes. Dans ce panorama, est également citée l'approche « distributionnelle » (Li, 2006), que nous adapterons par la suite.

Généralement, toutes les méthodes s'accordent pour ne pas sur corriger notamment les dates, montants et plus généralement les chiffres (numéro de téléphone, heure de rendez-vous, etc.), ce qui peut avoir de graves conséquences.

4 Présentation des méthodes

Notre approche s'inspire des travaux utilisant l'analyse distributionnelle, généralement mise en œuvre pour détecter des relations sémantiques comme la synonymie à l'aide de corpus textuels (Bourigault, 2002) et (Greffenstette, 1994). Nous reprenons cette approche mais pour détecter les variantes orthographiques des mots en comparant la distribution de leurs contextes. Ensuite, à partir du graphe des voisins orthographiques et contextuels, nous proposons deux méthodes pour corriger les textes bruités : l'une semi-automatique et l'autre complètement automatique.

Pour ce faire, nous nous sommes basés sur le fait qu'une forme bien orthographiée apparaît plus souvent dans le corpus que ses formes mal orthographiées et qu'ainsi, les contextes d'une forme mal orthographiée se retrouvent parmi ceux de la forme bien orthographiée. Par exemple, dans le « corpus 100k », on trouve 292 294 occurrences pour « client » et 220 657 pour « cliente », les formes mal orthographiées associées à ces mots ayant le plus d'occurrence étant « clt » apparaissant 87 257 fois et « clte » qui apparait 63 439 fois, jusqu'à 221 fois pour « clietn ». On peut expliquer ce phénomène parce que les fautes se répartissent sur un grand nombre de formes (une cinquantaine pour « client »).

Néanmoins, ceci n'est pas toujours vrai. En effet, pour certains types d'erreurs, en particulier pour la suppression de caractères accentués, les formes mal orthographiées sont aussi nombreuses voire plus nombreuses que les formes bien orthographiées. Dans le « corpus 100k », « recu » et « reçu » sont presque aussi fréquents (34 853 contre 36 999), et « chèque » compte 11 870 occurrences alors que la forme sans accents « cheque » apparait 15 592 fois.

La suite présente les deux méthodes après la partie préliminaire, commune aux deux méthodes.

4.1 Partie préliminaire commune aux deux méthodes

Pour résumer, cette partie cherche à établir un graphe constitué des mots ou formes qui se « ressemblent » et qui partagent des contextes similaires.

Etape 1 : pour tous les commentaires ou textes du « corpus 100k », la ponctuation est enlevée car elle n'est pas toujours mise à bon escient. Le texte est considéré comme une suite de mots m_i. Pour chaque mot m_i, les contextes sont calculés à l'aide des mots qui le précèdent et qui lui succèdent. Les formes sont prises de manière brutes sans analyse morpho-syntaxique. Pour chaque mot m_i (sauf pour les premiers et derniers), on obtient :

- 2 contextes simples (bigrammes) : « m_{i-1} _ », « _ m_{i+1} »
- 3 contextes doubles (trigrammes) : « m_{i-2} m_{i-1} _ », « m_{i-1} _ m_{i+1} », « _ m_{i+1} m_{i+2} »

L'association (« mot », « contexte ») est stockée dans une base de données Lucene[1], ce qui permet ensuite de trouver rapidement tous les contextes pour un mot donné ou de trouver les mots associés à un contexte donné.

[1] - Moteur de recherche développé par la fondation Apache (http://lucene.apache.org/)

Etape 2 : la base de données est parcourue afin d'éliminer les contextes uniques car pouvant amener du bruit. Pour le « corpus 100k », le nombre total de contextes est de 22 millions. La liste des formes présentes dans le corpus est établie et classée par ordre de fréquence décroissante.

Etape 3 : les formes de la liste précédente vont être comparées deux à deux à l'aide de deux mesures de similarité : $sim_{Damerau}$ et $sim_{Raccourcie}$. La mesure $sim_{Damerau}$ est basée sur la distance de Damerau-Levenstein (Damerau, 1964) qui consiste à calculer le nombre minimum d'opérations nécessaires pour transformer une chaîne de caractères en une autre, où une opération est définie comme l'insertion, la suppression, la substitution d'un simple caractère, ou encore la transposition de deux caractères. La valeur obtenue est divisée par le maximum des longueurs des deux chaînes à comparer. Pour « client » et « cliemnt » on obtient une distance de 0,1428 ou similarité de 0,8571.

Néanmoins cette mesure ne permet pas de trouver les formes raccourcies ou abrégées que l'on rencontre assez fréquemment comme « inter » pour « intervention » ou « cl » pour « client ». On pourrait quand même les trouver avec cette mesure en baissant le seuil limite mais au risque d'introduire du bruit. Ces réflexions nous ont amenés à imaginer la mesure $sim_{Raccourcie}$ qui consiste à compter le nombre de paires de lettres qui se suivent dans la chaîne de caractères la plus courte et qui font partie de la chaîne la plus longue, divisée par le nombre de paires de lettres qui se suivent de la chaîne la plus courte. On obtient ainsi un score de similarité de 1 entre « cl » et « client » ou entre « inter » et « intervention », mais, par exemple, un score de 0,5 entre « tenir » et « intervention ».

A l'aide d'une de ces deux mesures et d'un seuil ($seuil_{mot}$), la méthode permet de sélectionner des paires de mots candidats.

Etape 4 : les contextes vont ensuite permettre de déterminer si les deux mots candidats seront considérés comme des variations orthographiques ou non. Comme le nombre de contextes des mots peut varier fortement, il faut donc rester prudent sur le mode de comparaison. Nous adaptons une des mesures de (Bourigault, 2002) et calculons le ratio entre le nombre de contextes communs des deux mots et le nombre total de contextes du mot le moins fréquent :

$$ratio = \frac{\left| C_{mot\ +fr\acute{e}quent} \cap C_{mot\ -fr\acute{e}quent} \right|}{\left| C_{mot\ -fr\acute{e}quent} \right|}$$

Si ce ratio est supérieur à un seuil ($seuil_{contexte}$), on considère qu'un lien existe entre $mot_{\text{-fréquent}}$ et $mot_{\text{+ fréquent}}$.

Dans « corpus 100k », le mot « client » possède au total 260 703 contextes (dont 13 974 différents), « cleint » 235 contextes (dont 65 différents). « cleint » partage 224 contextes avec « client » (sur 235 au total), soit un ratio de 0,95, d'où la présence d'un lien entre « client » et « cleint ».

Au final, on obtient :

– Une liste des mots qui n'ont pas de variation orthographique, soit parce qu'ils n'en ont effectivement pas, soit parce que la méthode des contextes n'a pas

réussi à leur trouver des mots voisins.
– Un graphe de mots similaires orthographiquement et contextuellement.

A titre d'exemple, voici ce que peut donner une toute petite partie du graphe visualisé avec le logiciel Gephi[2], centré sur le mot « prélèvement » :

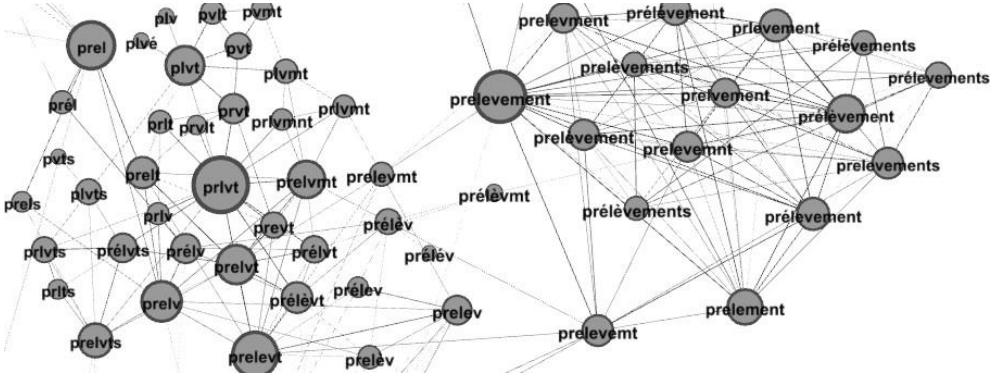

FIGURE 1 – Graphe des mots voisins du mot « prélèvement »

4.2 Méthode 1 : semi-automatique et dictionnaire

Cette méthode consiste à utiliser le graphe précédent pour générer un dictionnaire de substitution en détectant les parties connexes du graphe ou en appliquant un algorithme de détection de communauté comme (Blondel, 2008). Au final, on obtient des groupes de mots que l'on présente à l'expert en les classant par ordre de fréquences décroissantes, ce qui va lui permettre de modifier le dictionnaire de substitution de manière experte en fonction des connaissances qu'il a du domaine et des spécificités des données. En faisant varier $seuil_{mot}$ et $seuil_{contexte}$ de manière experte, il fera apparaître plus ou moins de mots et de relations entre les mots.

Une fois ce dictionnaire de substitution élaboré, les textes à corriger sont parcourus, caractère par caractère, et chaque fois qu'une suite de mots correspond à une entrée du dictionnaire, elle est remplacée par sa forme canonique.

4.3 Méthode 2 : vers le tout automatique

Cette autre méthode est beaucoup plus ambitieuse puisqu'elle cherche à corriger, automatiquement, les formes erronées en s'appuyant sur le réseau précédent. Pour corriger une phrase comme : « le cliemnt veut changer d'abonnmnt», elle commence par supprimer la ponctuation puis trouver les mots « candidats » à la substitution pour chaque mot de la phrase :

– « le », « veut », « changer » et « d » font partie de la liste des mots qui n'ont pas de variations ou appartiennent à une « stop liste », ils ne sont donc pas traités.
– « cliemnt » et « abonnment » font partie du réseau. Leurs substituants possibles

[2] - Logiciel de manipulation, d'édition et de visualisation de graphes (http://gephi.org/)

sont calculés à partir du réseau : il s'agit des pères et fils de ces mots.

Au final, on obtient :

le	cliemnt	veut	changer	d	abonnment
	client				abonnement
	cliente				
	…				

TABLE 2 – Liste des mots candidats à la substitution

L'étape suivante consiste à trouver, parmi les mots candidats, ceux qui vont maximiser la probabilité de rencontrer la phrase M, composée des mots m_i, selon la formule suivante (avec un lissage additif encore appelé « ajouter un » pour calculer la probabilité de rencontrer m_i sachant m_{i-2} et m_{i-1}(Beaufort, 2002), $|V|$ étant la taille du vocabulaire) :

$$P(M) = \prod_i P(m_i|m_{i-2}m_{i-1}) \quad avec \quad P(m_i|m_{i-2}m_{i-1}) = \frac{1 + nb(m_{i-2}, m_{i-1,}m_i)}{|V| + nb(m_{i-2}, m_{i-1},*)}$$

Comme le décrit (Cucerzan, 2004), il s'agit d'un problème d'optimisation locale : on calcule si le fait de changer « cliemnt » en « client » augmente la probabilité de l'ensemble. Ainsi, de manière itérative, on corrige la phrase. On peut ensuite lancer récursivement plusieurs corrections de la phrase, puisque le fait de corriger un mot va modifier le contexte de ses mots voisins et peut-être ainsi permettre des corrections lors des itérations suivantes. Nous avons observé ce phénomène, par exemple « recla » est corrigé en « reclamation » lors de la 1ère correction, puis en « réclamation » lors de la 2ème correction.

5 Résultats

Tous ces travaux se placent dans un contexte industriel. Il est donc nécessaire que les calculs puissent se faire dans des temps « raisonnables ». Ce point est acquis puisque le calcul du réseau de voisins sur le « corpus 100k » est obtenu en quelques dizaines de minutes sur un PC portable de moyenne gamme.

Pour ce qui est de la génération du dictionnaire de substitution à partir du graphe des voisins (méthode 1), les premiers résultats montrent que le fait de pouvoir générer une première version de ce dictionnaire que l'expert peut ensuite modifier à la main est appréciable, notamment pour assurer une large couverture des mots à corriger. Présenter ces groupes de mots triés par nombre d'occurrences totales permet à l'expert de se concentrer sur les formes erronées les plus importantes. Cette démarche a permis à l'équipe EDF Commerce de détecter des mots, abréviations, formulations ou raccourcis qui n'étaient pas pris en compte dans le processus actuel de correction.

Pour la méthode 2, entièrement automatisée, les résultats doivent être améliorés, notamment sur la manière de fixer les seuils. Cette méthode produit des erreurs :

– Sur les mots qui se ressemblent et qui partagent des contextes voisins comme « peut/veut », « semestrielle/bimestrielle » (employé dans « facture semestielle|bimestrielle ») ou encore « satisfait/insatisfait » (« client

satisfait|insatisfait »).
– Sur les mots dont l'orthographe erronée est aussi fréquente voire plus fréquente que l'orthographe correcte comme pour les mots « recu » ou « cheque ».

6 Conclusion et perspectives

Nous avons présenté deux méthodes, l'une semi-automatique, l'autre entièrement automatisée, pour la correction de réclamations rédigées par des conseillers. En construisant pour chaque mot un graphe des voisins orthographiques et contextuels, nous avons montré comment détecter ses formes mal orthographiées afin de construire un dictionnaire de substitution. En utilisant celui-ci dans la première méthode semi-automatique, nous avons amélioré le processus de normalisation des réclamations déjà existant. En outre, la deuxième méthode entièrement automatisée basée elle aussi sur les contextes, semble intéressante mais nécessite, du fait de la sur-correction, une grande vigilance. Néanmoins, ces travaux ne sont pas terminés et constituent le début de développements et de tests notamment par le biais d'un corpus d'évaluation en cours d'élaboration.

Références

BARANES, M. (2012). Vers la correction automatique de textes bruités : Architecture générale et détermination de la langue d'un mot inconnu. In *RECITAL'2012-Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*.

BEAUFORT, R., DUTOIT, T., & PAGEL, V. (2002). Analyse syntaxique du français. Pondération par trigrammes lissés et classes d'ambiguïtés lexicales. Proc. JEP, 133-136.

BLONDEL, V. D., GUILLAUME, J. L., LAMBIOTTE, R., & LEFEBVRE, E. (2008). Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008(10), P10008.

BOURAOUI, J. L., BOISSIÈRE, P., MOJAHID, M., VIGOUROUX, N., LAGARRIGUE, A., VELLA, F., & NESPOULOUS, J. L. (2009). Problématique d'analyse et de modélisation des erreurs en production écrite. Approche interdisciplinaire. *Actes de TALNRECITAL 2009*.

BOURIGAULT, D. (2002, June). Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. In *Actes de la 9ème conférence annuelle sur le Traitement Automatique des Langues (TALN 2002), Nancy* (pp. 75-84).

CUCERZAN, S., & BRILL, E. (2004, July). Spelling correction as an iterative process that exploits the collective knowledge of web users. In *Proceedings of EMNLP* (Vol. 4).

DAMERAU, F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM*, *7*(3), 171-176.

GREFENSTETTE, G. (1994). *Explorations in automatic thesaurus discovery*. Springer.

LI, M., ZHANG, Y., ZHU, M., & ZHOU, M. (2006, July). Exploring distributional similarity based models for query spelling correction. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics* (pp. 1025-1032). Association for Computational Linguistics.

