Inbenta Semantic Clustering : un outil de classification non-supervisée hybride

Quintana Manon, Planes Laurie
Inbenta France, 164 route de Revel, 31400 Toulouse, France
mquintana@inbenta.com, lplanes@inbenta.com

RÉSUMÉ

Inbenta développe un outil de classification non-supervisée hybride qui allie à la fois les statistiques et la puissance de notre lexique inspiré de la Théorie Sens-Texte. Nous présenterons ici le contexte qui a amené à la nécessité de développer un tel outil. Après un rapide état de l'art sur la classification non-supervisée en TAL, nous décrirons le fonctionnement de notre clustering sémantique.

ABSTRACT

Inbenta Semantic Clustering : a hybrid unsupervised classification tool
Inbenta develops a hybrid usupervised classification tool combining both statistics and the power of our Meaning-Text Theory-based lexicon. We present here the context which lead us to develop such a tool. After a quick state of the art of unsupervised classification (clustering) in NLP, we will describe how our semantic cluster works.

MOTS-CLÉS : classification non-supervisée, sémantique, théorie Sens-Texte, fonctions lexicales, statistique, K-medoïdes

KEYWORDS: unsupervised clustering, semantics, meaning-text theory, lexical functions, statistics, K-medoids

1 Contexte

1.1 Notre moteur de recherche sémantique

Les entreprises font face à une forte exigence de leurs clients en termes de service. Ceux-ci souhaitent obtenir les réponses à leurs questions facilement et rapidement. Or, les canaux traditionnels de contact (téléphone et mail) coûtent cher à ces entreprises car ils nécessitent la mise en place de moyens humains et infrastructurels importants. Aussi, avec la culture numérique, les usagers ont l'habitude de rechercher des informations via Internet et d'y accéder de façon immédiate et autonome, ce qui n'est pas le cas d'un appel téléphonique ou d'un email. De plus, dans la grande majorité des cas, ces recherches sont relativement basiques et fortement redondantes. L'intervention d'un interlocuteur n'est pas nécessaire, une FAQ ou une page du site suffiraient à y répondre. Aujourd'hui, les entreprises doivent donc faire face à ces nouvelles habitudes et proposer de nouveaux outils à leurs clients afin d'améliorer l'expérience de la relation client.

C'est dans ce contexte que nous avons développé un moteur de recherche intelligent. Ce moteur est intégré sur les sites d'entreprises. Il facilite l'accès à l'information et ce, en trouvant la réponse dans la base de connaissances qui correspond le mieux à la requête de l'internaute posée en langage naturel.

Notre moteur de recherche sémantique est intégré chez une trentaine de clients français, la majorité appartenant au secteur de la banque et de l'assurance.

1.2 Pourquoi le Clustering Sémantique ?

Pour améliorer la qualité de son moteur de recherche sémantique, nous proposons aux entreprises d'enrichir leurs bases de connaissances en fonction des demandes des internautes. En effet, en fonction des nouvelles lois, des offres commerciales ou des thématiques du moment, la base de réponses nécessite d'être optimisée et complétée.

Pour cela, il faut analyser toutes les questions utilisateurs qui n'ont pas obtenu de réponse et ajouter des contenus dans la base de connaissances selon ces analyses. Il s'agit d'un travail conséquent en raison des centaines de milliers de requêtes à analyser. Nous avons donc développé un outil qui analyse automatiquement ces requêtes et les classe par thématiques. Cet outil de classification non-supervisée, appelé Inbenta Semantic Clustering (ISC), permet ainsi de réduire grandement ce temps d'analyse.

2 Etat de l'art de la classification non-supervisée en TAL

La classification automatique de texte a pour but de regrouper automatiquement dans un même ensemble (ou classe) des textes thématiquement proches.

Deux types de classification sont distingués. Dans l'approche dite « supervisée », les classes sont définies au préalable par un expert tandis que dans une approche dite « non supervisée », les classes émergent des calculs de la machine. Comme indiqué précédemment, on s'intéressera spécifiquement au dernier cas, aussi appelé clustering.

Dans cette tâche, les données des groupes (ou clusters) sont considérées comme proches lorsque les données appartenant à un même cluster sont les plus similaires possibles, et les données de groupes différents sont les plus « dissemblables » possibles

Les grandes étapes du clustering sont : la définition des variables, le choix des mesures de distances, le choix de l'algorithme de regroupement.

2.1 Variables

Dans le domaine du clustering de données textuelles, la première étape nécessite de choisir une représentation des documents. Il faut donc se demander quels descripteurs seront utilisés. Parmi les formes de surface on compte les n-grammes et les sacs de mots. D'autres descripteurs tiennent compte des variations morphologiques (genre, nombre, conjugaison) comme les stems et les lemmes.

Un autre type d'approche s'intéresse au sens des documents et prend en compte la similarité sémantique. Cette dernière peut porter sur un niveau lexical (relations sémantiques entre les termes) ou sur un niveau conceptuel qui fait appel à des représentations conceptuelles telles que les ontologies. Nous nous concentrerons sur le niveau lexical.

2.2 Mesure des distances

(Wang et al. 2013) distinguent deux approches de la similarité sémantique au niveau lexical : « la première rassemble les mesures fondées sur des connaissances élaborées manuellement prenant typiquement la forme de réseaux lexicaux de type WordNet ; la seconde recouvre les mesures de nature distributionnelle, construites à partir de corpus. »

Dans l'étude de (Wang et al. 2013), les méthodes distributionnelles obtiennent de meilleurs résultats que les méthodes faisant appel à des connaissances. Il faut ici questionner la qualité de la ressource et son adéquation aux documents utilisés dans l'expérience : quelle est sa couverture par rapport au domaine des documents ? Comment sont résolues les ambiguïtés ?

On comprend facilement le choix du recours aux méthodes distributionnelles qui présentent l'avantage de ne pas nécessiter de ressource lexicale a priori, seulement de disposer d'un corpus conséquent sur le domaine étudié. Aussi, elles sont adaptables à de nouveaux domaines et à d'autres langues. Nous verrons par la suite que notre approche évacue cette question car nous possédons déjà une ressource, qui est à la base de nos travaux et s'enrichit de façon continue.

2.3 Algorithmes

Concernant l'algorithme de clustering nous retiendrons les deux types d'approches principaux : les méthodes hiérarchiques et les méthodes par partitionnement.

Les méthodes hiérarchiques, (Classification hiérarchique ascendante CAH, Classification hiérarchique descendante, CURE, BIRCH, l'algorithme de Ward, etc.) procèdent par agglomérations ou divisions successives. Pour les hiérarchies ascendantes on part avec un élément par classe, puis à chaque étape on fusionne les deux classes les plus proches. Pour les hiérarchies descendantes le processus est inversé (on part avec tous les éléments dans une seule et même classe).

Les méthodes de clustering par partitionnement (K-Means, Fuzzy C-Means, IsoData, Fast Global K-Means, K-Means++) « proposent, en sortie, une partition de l'espace des objets (…). Le principe est alors de comparer plusieurs schémas de clustering (plusieurs partitionnements) afin de retenir le schéma qui optimise un critère de qualité (…) en procédant de façon itérative, en améliorant un schéma initial choisi plus ou moins aléatoirement, par ré-allocation des objets autour de centres mobiles. » (Cleuziou, 2004). Notre outil de clustering appartient à cette seconde famille de méthodes.

3 Fonctionnement de l'Inbenta Semantic Clustering (ISC)

L'objectif de l'ISC est de détecter de nouvelles thématiques de requêtes. Nous n'avons donc pas de classes prédéfinies, ce qui nous place dans le cadre d'une classification non supervisée.

La tâche de détection de thématiques nécessite de mesurer la similarité sémantique des requêtes. Or, nous possédons une ressource lexicale construite spécialement pour le moteur de recherche dont les requêtes à classer sont issues. C'est pourquoi nous avons naturellement décidé d'exploiter cette ressource pour mesurer la similarité sémantique des requêtes.

La ressource est notre lexique construit pour faire de l'expansion de requête dans le cadre d'un moteur de recherche sémantique. Il s'inspire de la théorie Sens-Texte (Melcuk, 1995) dans la mesure où il décrit les termes par les relations sémantiques (Fonctions Lexicales) qu'ils entretiennent.

La théorie Sens-Texte (TST) propose un modèle du langage dit « traductif » qui permet de mettre en correspondance des représentations sémantiques avec toutes les représentations phoniques qui peuvent les exprimer dans une langue donnée. Il s'agit donc d'un modèle reposant sur le paraphrasage.

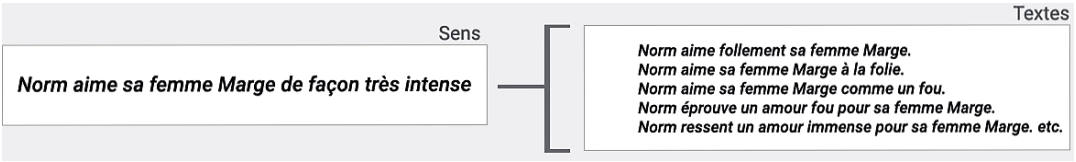

Figure 1: La TST, un modèle traductif

Les relations sémantiques entre termes sont matérialisées par des Fonctions Lexicales : « La vocation des fonctions lexicales est de fournir au locuteur la totalité des moyens lexicaux nécessaires à l'expression la plus riche, la plus variée et la plus complète de la pensée et, en même temps, de garantir le choix le plus précis de la formulation appropriée. En d'autres termes, les FL […] alimentent un système puissant de paraphrasage, qui est à la fois une raison d'être des FL et un outil fondamental de leur vérification.» (Mel'čuk, 1995)

Notre lexique vise la description limitée du domaine de la banque et de l'assurance, dans la limite des mots effectivement employés dans nos projets afin d'éviter de générer des ambiguïtés qui n'ont pas lieu d'être au sein d'un domaine spécifique. Voici comment il s'insère dans la chaîne de traitement du clustering sémantique:

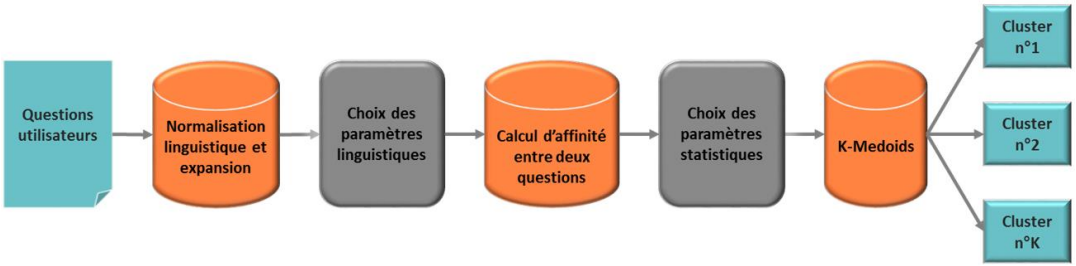

Figure 2 : Schéma de fonctionnement de l'ISC

Tout d'abord, nous normalisons les requêtes (questions utilisateurs) via notre correcteur orthographique. Ensuite nous lemmatisons les termes en cherchant le lemme correspondant au mot dans le lexique, nous recourons à des règles morphosyntaxiques de désambiguïsation. Ainsi, nous décrivons chaque requête à classer par un vecteur de lemmes : les lemmes effectivement présents dans la requête ainsi que les lemmes sémantiquement reliés (ie. présents dans les fonctions lexicales). Il s'agit donc d'une forme d'expansion sémantique des requêtes.

Les lemmes sont pondérés en fonction des poids qui leurs sont attribués dans notre lexique sur la base de leur catégorie grammaticale (les déterminants, par exemple, ont un poids nul) et de leur importance sur le site. Sur maaf.fr, le lemme « Maaf » aura un poids faible car l'évocation de Maaf sur son propre site est beaucoup moins chargée de sens que sur un site concurrent.

Les lemmes « étendus » sont pondérés en fonction de leur proximité avec les lemmes effectivement présents. En effet, nous avons préalablement défini dans le lexique un poids de proximité par type de relation. Par exemple, le participe passé associé à un verbe a un poids de proximité de 0,9/1 par rapport au verbe.

Nous choisissons ensuite quelles catégories de lemmes nous prenons en compte (catégories grammaticales et poids sémantiques) ainsi que les relations conservées pour l'expansion.

A ce stade, nous calculons l'affinité entre chaque couple de questions étendues.

Nous pouvons alors choisir une configuration statistique pour le clustering : nombre de clusters, nombre de candidats centroïdes utilisés, nombre d'itérations. Les centroïdes sont les centres de gravité des clusters, ce sont les points les plus centraux, à partir desquels les éléments sont rassemblés.

Ensuite, nous lançons l'algorithme K-Medoïdes, moins sensible que K-means aux points aberrants (Park et al., 2006), qui effectue un partitionnement des requêtes en K clusters (K étant défini au préalable). Pour cela, il effectue plusieurs itérations qui tentent de minimiser la distance entre les éléments de la partition. Nous avons choisi cet algorithme car il est simple à mettre en œuvre. Nous l'avons testé en concurrence avec l'algorithme « Affinity Propagation » qui présentait l'avantage de ne pas devoir définir préalablement le nombre de partitions. Contrairement au K-Medoïdes, qui considère tour à tour plusieurs éléments comme candidats centroïdes, Affinity Propagation « considère simultanément l'ensemble des points comme des centroïdes, qui vont eux-mêmes échanger des messages pour déterminer lesquels sont les meilleurs candidats. » (Liu, 2016). Etant donné que le K-Medoïdes obtenait de meilleures performances, nous avons décidé de nous concentrer sur ce dernier.

Nous obtenons en sortie des groupes de requêtes sémantiquement proches.

Les différents clusters sont hiérarchisés par combinaison du score d'affinité moyen et de la représentativité du cluster au sein du corpus. Ainsi, sont présentés en premier les clusters les plus fiables et qui rassemblent les plus grandes parties du corpus. Ceci permet de reléguer en dernière position les grands clusters dont les éléments sont plutôt éloignés, de même que les clusters contenant peu d'éléments mais qui sont très proches les uns des autres. Nous utilisons actuellement les centroïdes comme titres des clusters.

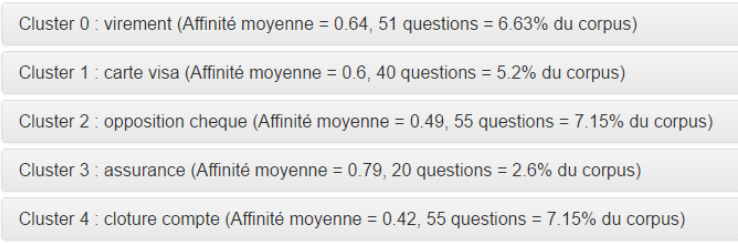

Figure 3 : Tri des clusters

Au sein d'un cluster, les requêtes sont triées par affinité décroissante avec le centroïde. Ceci nous permet de voir les requêtes les plus proches en premier lieu et également de repérer à partir de quel seuil d'affinité la qualité du cluster se dégrade.

Cluster 0 : Devis assurance (Affinité moyenne = 0.9, 341 questions = 20.73% du corpus)

Rang x nb d'apparitions	Question	Affinité
0 x1	Devis assurence	100%
1 x12	Devis assurance	100%
2 x1	devi	100%
3 x3	devis assurance	100%
4 x151	Devis	100%
5 x3	Dzvis	100%
6 x2	simulation	80%
7 x2	Simulation assurance	80%
8 x2	Souscrire	80%
9 x21	Simulation	80%

Figure 4 : Tri des requêtes au sein d'un cluster

4 Conclusions et perspectives

Notre outil de clustering sémantique est toujours en cours de développement. Nous en sommes à l'étape de consolidation du prototype. Plusieurs points sont à améliorer comme par exemple la lisibilité des titres de cluster. En effet, ceux-ci ne sont autres que les éléments du corpus d'entrée choisis comme centroïdes de chaque cluster. Il s'agit donc des données brutes laissées par les utilisateurs et celles-ci peuvent contenir des erreurs de syntaxe ou d'orthographe. Une de nos évolutions est donc de standardiser ces centroïdes.

Une autre des perspectives à mettre en place est la configuration des paramètres linguistiques et statiques évoqués plus haut. En effet, ces paramètres seront différents selon la taille du corpus. Nous sommes actuellement en cours d'évaluation de l'outil afin de déterminer les paramètres optimums.

Cette évaluation est également un gros travail qu'il nous reste à effectuer. Cependant, nous avons constaté lors de notre état de l'art que l'évaluation d'un clustering textuel non-supervisé n'est pas chose aisée. En l'absence de classification de référence, nous devons nous « tourner vers l'exploitation d'indices de qualité utilisables en mode non supervisé. Un état de l'art du domaine permet de recenser des indices basés sur des calculs de distance (...) [Des] expérimentations ont cependant montré qu'aucun de ces indices ne permettait d'estimer correctement la qualité d'un résultat de clustering sur des données textuelles. Ceux-ci ne permettent notamment pas de discriminer entre des résultats de classification homogènes et des résultats hétérogènes. Ils peuvent même présenter le défaut important de privilégier cette dernière famille de résultats » (Cuxac et al. 2010)

Références

CLEUZIOU G. (2004) Une méthode de classification non-supervisée pour l'apprentissage de règles et la recherche d'information. Université d'Orléans : autre

CUXAC P., LAMIREL J.C., GHRIBI M. (2010) Les méthodes de classification non supervisées appliquées aux textes : mesure de la performance des résultats de clustering de documents.

PARK H.S., LEE J.S., JUN C.H., (2006) A K-means-like Algorithm for K-medoids Clustering and Its Performance," ICCIE 2006, 102-117

MEL'CUK I., POLGUERE A. (1995) *Introduction à la lexicologie explicative et combinatoire*. Bruxelles : Duculot.

WANG W., BESANCON R., FERRET O., GRAU B. (2013) Regroupement sémantique de relations pour l'extraction d'information non supervisée. TALN-RÉCITAL 2013, 353-366

LIU Y., LI J., PLAZA A. (2016) Spectrometer-Driven Spectral Partitioning for Hyperspectral Image Classification. *IEEE Journal of selected topics in applied earth observations and remote sensing*, VOL. 9, 668-680