
RÉSEAUX DE NEURONES POUR LE TRAITEMENT AUTOMATIQUE DU LANGAGE

CONCEPTION ET RÉALISATION DE FILTRES D'INFORMATIONS

Mathieu STRICKER

ESPCI, Laboratoire d'Électronique – Université Paris 6
10, rue Vauquelin, F - 75005 Paris FRANCE
Tél. : +33 1 40 79 46 95
Mel : Mathieu.Stricker@espci.fr

En raison de l'augmentation constante du volume d'information accessible électroniquement, la conception et la mise en œuvre d'outils efficaces, permettant notamment à l'utilisateur de n'avoir accès qu'à l'information qu'il juge pertinente, devient une nécessité absolue. Comme la plupart de ces outils sont destinés à être utilisés dans un cadre professionnel, les exigences de fiabilité et de convivialité sont très importantes ; les problèmes à résoudre pour satisfaire ces exigences sont nombreux et difficiles. L'accès à cette information pertinente peut se faire en fournissant à un utilisateur des documents pertinents ou en lui proposant des passages de documents pertinents (ou des réponses à des questions). Le premier cas relève du domaine de la *recherche de textes* et le second du domaine de l'*extraction d'informations*.

C'est dans le domaine très actif de la recherche de textes que s'est situé notre travail, réalisé dans le cadre d'une collaboration entre Informatique CDC, filiale de la Caisse des Dépôts et Consignations, et le Laboratoire d'Électronique de l'ESPCI.

Le but de nos travaux a été de développer un modèle fondé sur l'apprentissage numérique pour la catégorisation de textes ou, plus précisément, pour ce qui correspond à la tâche de *routing* dans le découpage de la conférence TREC (*Text REtrieval Conference*). L'approche que nous avons conçue nous a permis d'obtenir un résultat très satisfaisant : nous avons remporté la tâche de *routing* de la compétition TREC 9, devançant notamment Microsoft.

Le point essentiel de notre approche est l'utilisation d'un classifieur qui est un réseau de neurones dont l'architecture prend en considération le contexte local des mots. La mise en œuvre d'une méthode de sélection des entrées nous a permis de réduire à une vingtaine le nombre de descripteurs de chaque texte ; néanmoins, le nombre de paramètres reste élevé eu égard au nombre d'exemples disponibles (notamment lors de la compétition TREC 9). Il a donc été nécessaire de mettre en œuvre une méthode de régularisation pour obtenir des résultats significatifs à l'issue des apprentissages.

Nos résultats ont été validés d'une part grâce au corpus Reuters-21578⁸ qui est souvent utilisé par la communauté de la catégorisation de textes, et d'autre part, par la participation aux sous-tâches de *routing* de TREC-8 et TREC-9, qui ont permis d'effectuer des comparaisons chiffrées avec d'autres approches.

Nos travaux ont été intégrés dans l'application ExoWeb développée à la Caisse des Dépôts, pour y ajouter des fonctionnalités opérationnelles originales. Cette application offrait, sur l'intranet du groupe, un service de catégorisation de dépêches AFP en temps réel ; cette catégorisation s'effectuait grâce à des modèles à bases de règles.

La première fonctionnalité nouvelle résultant de nos travaux est un outil qui permet à l'administrateur du système de surveiller automatiquement le vieillissement de filtres construits sur des modèles à base de règles. L'idée de cette application est de fabriquer une "copie" d'un filtre à base de règles avec un filtre utilisant un réseau de neurones. Comme le réseau de neurones produit une probabilité de pertinence et non une réponse binaire, il est possible d'attirer l'attention de l'administrateur sur les documents pour lesquels les filtres et les réseaux de neurones fournissent des réponses divergentes : documents considérés comme pertinents par la méthode à base de règles, mais obtenant une probabilité proche de zéro avec le réseau de neurones, et documents considérés comme non pertinents avec le premier et obtenant une probabilité de pertinence proche de un avec le second.

Nous avons également proposé les bases d'une deuxième application, qui permet à un utilisateur de fabriquer lui-même un filtre à sa convenance avec un travail minimum. Pour réaliser cette application, il est nécessaire que l'utilisateur fournisse une base de documents pertinents. Cela peut se faire grâce à l'utilisation d'un moteur de recherche conjointement avec un réseau de neurones ou uniquement grâce au moteur de recherche.

Le chapitre 1 est une introduction à la problématique abordée. Le chapitre 2 est une présentation des modèles couramment utilisés en recherche d'informations, comme le modèle vectoriel ou le modèle probabiliste, qui sont à l'origine de beaucoup de modèles construits pour la catégorisation de textes.

Le chapitre 3 présente les corpus utilisés tout au long de cette étude. Les spécificités de chacun de ces corpus sont soulignées. Ces corpus sont disponibles gratuitement et ont été utilisés par d'autres auteurs, ce qui facilite les comparaisons.

⁸ Ce corpus est publiquement accessible sur le site : <http://www.att.research.com/~lewis/reuters21578.html>

Le chapitre 4 introduit les différentes mesures utilisées pour évaluer les performances des systèmes. Au-delà des définitions, ce chapitre met en évidence le bruit inhérent à ces mesures ; à l'heure actuelle, les performances absolues n'ont pas un grand sens : seules les performances relatives sont importantes.

Le chapitre 5 montre comment les textes sont transformés pour pouvoir être utilisés par les méthodes d'apprentissage numérique. Ce chapitre met en évidence la nécessité et la difficulté d'effectuer une sélection de descripteurs. Nous proposons une méthode entièrement automatique en deux étapes. La première étape détermine le vocabulaire spécifique des documents pertinents par rapport à l'ensemble du corpus ; la deuxième étape est une procédure d'orthogonalisation selon la méthode de Gram-Schmidt qui présente l'avantage d'être adaptée à la classification.

Le chapitre 6 est une présentation succincte des réseaux de neurones. Ce chapitre insiste sur la notion de surapprentissage pour les problèmes de classification, et montre que les méthodes de régularisation comme le *weight decay* apportent une solution à ce problème tout en ajoutant de nouveaux paramètres appelés hyperparamètres. Ces hyperparamètres peuvent être théoriquement déterminés grâce à l'approche bayésienne qui est présentée avec les approximations nécessaires à sa mise en œuvre.

Le chapitre 7 présente les premières expériences effectuées sur le corpus Reuters ainsi que la description de notre participation à TREC-8. Ce chapitre étudie l'impact des différents paramètres intervenant dans la sélection des descripteurs sur les performances (nombre de documents non pertinents, choix de ces documents, nombre de descripteurs initiaux). L'étude du nombre optimal de neurones cachés montre qu'il est nécessaire d'améliorer la représentation des textes avant de complexifier l'architecture des réseaux de neurones en ajoutant des neurones cachés.

Le chapitre 8 présente une méthode originale pour déterminer automatiquement le contexte caractéristique d'un mot pour effectuer une désambiguïsation partielle de ce mot. L'architecture neuronale est modifiée pour prendre en considération cette nouvelle représentation. Avec celle-ci, l'utilisation d'une méthode de régularisation est indispensable ; malheureusement les résultats de l'approche bayésienne n'ont pas permis d'obtenir des résultats satisfaisants pour la détermination des hyperparamètres. Finalement les résultats obtenus sur le corpus Reuters et sur le corpus de TREC-8, montrent une amélioration notable des résultats par rapport au chapitre 7. Ce chapitre se termine par la description de notre participation à TREC-9.

Enfin, le chapitre 9 montre comment ces résultats sont intégrés dans une application existante de filtrage de dépêches de l'AFP en temps réel. Les méthodes d'apprentissage numérique permettent de proposer de nouvelles fonctionnalités à cette application. Une première application utilise la sortie d'un filtre construit sur des méthodes à base de règles conjointement avec la sortie d'un filtre neuronal pour surveiller le premier filtre. Une deuxième application utilise un moteur de recherche couplé aux méthodes neuronales pour permettre à un utilisateur de définir son propre filtre.

AMORÇAGE PERCEPTIF VERBAL RÔLE DE LA CONGRUENCE SYLLABIQUE

Sandrine BELIER

Laboratoire d'Étude des Mécanismes Cognitifs - Université Lumière Lyon 2

Directeur de recherche : Professeur Olivier KOENIG

5, avenue Pierre Mendès France – F-69676 Bron cedex, France

Mel : sandrine.belier@univ-lyon2.fr

Mots clés : *Amorçage perceptif, recouvrement perceptif, identification perceptive, complètement de trigrammes, structure syllabique, compétition lexicale, asymétrie hémisphérique fonctionnelle.*

Au cours de ces quinze dernières années, la mémoire implicite a suscité un intérêt expérimental considérable dans le cadre des études en psychologie cognitive. Dans ce travail de thèse, nous nous sommes plus particulièrement intéressés à l'amorçage, phénomène de récupération implicite, automatique de l'information en mémoire. L'effet d'amorçage s'exprime par une influence, sur le traitement d'un stimulus (*cible*), résultant du traitement antérieur du même stimulus ou d'un stimulus lié au premier (*amorçe*).

Les principaux objectifs de ce travail étaient :

- 1) d'étudier le rôle de l'étendue du recouvrement perceptif entre amorçe et cible sur l'effet d'amorçage dans des tâches de complètement de bi- et trigrammes et d'identification perceptive (expériences 1, 2, 3, 6),
- 2) d'évaluer l'effet, sur l'amorçage, de la structure syllabique des mots utilisés dans ces mêmes tâches (expériences 1, 2, 4, 5, 6).

Ces tâches sont des outils utilisés dans le domaine de la neuropsychologie cognitive afin d'évaluer l'intégrité de la mémoire implicite chez des patients ayant des troubles de mémoire. Leur utilisation s'est avérée critique dans ce domaine, puisque ces tâches ont permis de mettre en évidence, dans les années 1980, une intégrité de la mémoire implicite chez les patients amnésiques. La tâche de complètement de trigrammes est celle dont nous nous sommes plus