UNIVERSITÉ DU QUÉBEC À MONTRÉAL

APPRENTISSAGE NON SUPERVISÉ DE LA SEGMENTATION LEXICALE AUTOMATIQUE DU CHINOIS BASÉ SUR LES RÉSEAUX BAYÉSIENS AVEC APPLICATION AUX TEXTES DES MÉDIAS SOCIAUX

THÈSE

PRÉSENTÉE

COMME EXIGENCE PARTIELLE

DU DOCTORAT EN INFORMATIQUE COGNITIVE

PAR

ZHE FU

MAI 2018

REMERCIEMENTS

J'aimerais tout d'abord remercier mon directeur de la recherche, monsieur Pierre Poirier, professeur à l'Université du Québec à Montréal, département de philosophie, qui m'a accompagné tout au long de ma thèse. Mes études au doctorat en informatique cognitive n'auraient jamais été terminées sans son encadrement, son conseil, son soutien, sa compréhension et sa patience. Sa disponibilité et ses généreux secours au cours de certains de mes moments difficiles ont été d'une très grande qualité et d'un immense réconfort ; merci infiniment Monsieur Pierre Poirier.

Je tiens à remercier ma codirectrice, Madame Fatiha Sadat, professeur à l'Université du Québec à mon Montréal, département d'informatique, pour les conseils sur mon projet de thèse.

Ma famille m'a donnée un grand support durant toutes ces années d'études doctorales au Canada, dont je ne pourrai mesurer l'apport que dans l'accomplissement de cette thèse. Les mots me manquent pour remercier, à sa juste valeur, ma femme, Wenna Liu, pour ses soutiens moraux et psychologiques indispensables pour maintenir ce projet à flot au travers des aléas de la vie, pour avoir cru en mes capacités intellectuelles et à mon sens de l'organisation pour le réaliser.

Encore un grand merci à tous pour m'avoir conduit à ce jour mémorable.

TABLES DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

RÉSUMÉ

La langue chinoise écrite présente une forme différente des langues alphabétiques latines. En chinois, une unité lexicale (un mot) peut contenir un ou plusieurs caractères et il n'existe pas d'espace entre les mots. Les lecteurs du chinois écrit doivent donc segmenter eux-mêmes la séquence de caractères et des segmentations différentes des caractères peuvent changer complètement la signification de la phrase. Il existe déjà plusieurs segmenteurs automatiques du chinois, mais ceux-ci, se basant sur un dictionnaire de mots connus, rencontrent souvent des difficultés lorsqu'ils sont confrontés à de nouveaux mots, à de nouvelles expressions et à de nouveaux symboles, ce qui peut réduire significativement leur performance. C'est en particulier le cas des textes tirés des médias sociaux (par exemple, Facebook, weibo, etc.), qui contiennent beaucoup de nouvelles expressions, de nouveaux mots, etc.

Nous proposons une méthode utilisant la segmentation du pinyin et les réseaux bayésiens, afin d'améliorer la performance des segmenteurs automatiques existants pour le domaine général et celui des médias sociaux. Cette méthode permet de calculer, d'évaluer et de mettre à jour automatiquement la probabilité qu'une séquence de caractères forme un mot. Aussi, lorsque la probabilité d'un nouveau mot, d'une expression ou d'un symbole s'avère élevée, cette méthode permet d'ajouter ce nouveau mot à un dictionnaire existant; c'est-à-dire, d'apprendre de nouveaux mots et d'ajuster continuellement la probabilité de chaque mot, y compris les nouveaux.

Appliquée de manière récursive, cette méthode peut améliorer la performance des segmenteurs en leur permettant d'apprendre de nouveaux mots et en mettant à jour automatiquement, et selon les différents corpus, les probabilités des mots de son dictionnaire. Cette méthode nous montre que la performance de la segmentation est

meilleure que celle des autres méthodes de segmentation, par exemple, le segmenteur Stanford[1], le NLPIR[2] et le LTP-Cloud[3] sur le corpus des médias sociaux et des journaux. Ainsi, sur un petit corpus de texte de journaux, notre système obtient une F-Mesure de 0,864, ce qui est supérieur à celle du segmenteur de Stanford à 0.768, celle de NLPIR à 0.789 et celle de LTP-Cloud à 0.811. De plus, sur un petit corpus de textes des médias sociaux, notre système obtient une F-Mesure de 0.907, ce qui est plus élevé que Stanford (0.611), NLPIR (0.693) et LTP-Cloud (0.793). Sur un grand corpus de textes de journaux, notre système obtient une F-Mesure de 0.924, ce qui est plus élevé que celle du segmenteur de Stanford (0.805), du NLPIR (0.759) et du LTP-Cloud (0.833). Enfin, sur un grand corpus des médias sociaux, notre système obtient une F-Mesure de 0.779, ce qui est plus élevé que Stanford (0.390), NLPIR (0.486) et LTP-Cloud (0.548).

Mots clés : Segmentation des mots chinois, Segmentation du pinyin, Réseau bayésien, Média social

[1] https://nlp.stanford.edu/software/segmenter.html

[2] http://nlpir.org/

[3] https://www.ltp-cloud.com/

INTRODUCTION

Cette thèse s'intéresse à la segmentation des textes, plus particulièrement ceux extraits des médias sociaux en langue chinoise. Notre objectif consiste à proposer une méthode permettant d'améliorer la performance de la segmentation des textes chinois issus aussi bien des journaux que des médias sociaux chinois. La force de ce travail réside dans la combinaison, par intermédiaire d'un réseau bayésien, de deux méthodes de segmentation : celle du texte en caractères chinois (sinogrammes ou hanzi) et celle du même texte en pinyin[4].

À la section 1.1, nous soulignons le fait que l'évolution nécessaire de la performance des procédés de segmentation répond à l'exigence de segmentation particulière aux textes chinois des médias sociaux. Sur ce plan, les outils actuels de segmentation présentent des avantages, mais aussi des points faibles. Selon l'évolution de la langue chinoise, ces points faibles réduisent de plus en plus la performance de la segmentation des textes chinois, et c'est dans le cas des textes tirés des médias sociaux que cette évolution est la plus manifeste. Ce problème est discuté à la sous-section 1.2. L'intégration du réseau bayésien est plus perfectionnée dans le cas de notre outil que pour les outils précédents (par exemple : le segmenteur de Stanford[5]), car elle considère tous les facteurs qui peuvent influencer la performance de la segmentation. Cette contribution est introduite à la section 1.3, où nous démontrons également les avantages de notre méthode de segmentation par rapport

[4] La version romanisée des caractères hanzi.

[5] https://nlp.stanford.edu/software/segmenter.shtml

aux autres méthodes en ce qui concerne les textes des médias sociaux. Enfin, l'organisation du rapport est détaillée à la section 1.4.

Généralités

Avec le développement des technologies, les médias sociaux sont de plus en plus utilisés, notamment depuis les années 2000. Aujourd'hui, tout le monde peut facilement publier, partager ou retrouver des informations grâce à ces médias. Afin que le texte soit plus court et plus clair, les plateformes Twitter, Facebook et Weibo ont limité le nombre des caractères publiés (140 caractères) et fixé d'autres contraintes spatiales ou temporelles. Certains médias sociaux ont, depuis, annulé cette limite, mais beaucoup de gens publient toujours leurs informations sous un format court, utilisant souvent une façon simple pour exprimer leurs idées ou des choses complexes. Ainsi ils peuvent utiliser des phrases incomplètes, voire simplement quelques mots, etc., pour décrire les informations. Les différentes utilisations des médias sociaux seront présentées à la section qui suit.

En général, les contenus de médias sociaux ne présentent pas un format standard, puisque chaque personne peut publier les informations à sa façon. C'est là une différence importante entre les textes que l'on retrouve dans les journaux et ceux tirés des médias sociaux. À peu de choses près, la grammaire des textes de journaux est toujours standard et les phrases ne contiennent pas ou très peu de symboles ou de pinyin. Cependant, les textes provenant des médias sociaux contiennent des phrases dont la grammaire n'est pas standard, ou encore des phrases qui contiennent de nouveaux mots, des émoticônes (symboles d'émotions), des symboles et des caractères chinois, etc. En effet, dans la langue chinoise, il y a deux systèmes d'écriture du langage : le hanzi et le pinyin (voir le chapitre II). Sur les médias sociaux, il arrive que les gens utilisent les deux systèmes d'écritures dans une même phrase.

Ces différences posent un grand défi pour les segmenteurs de textes chinois. En général, les segmenteurs peuvent bien segmenter les textes dont la grammaire est standard (par exemple, les textes de journaux), qui ne mélangent pas les symboles ou les alphabets, mais leur performance diminue lorsque les textes ne présentent pas ces caractéristiques, comme c'est le cas pour les textes de médias sociaux. En conséquence, le but de cette thèse est de développer un outil qui peut traiter ce genre de textes. Notre système peut aussi être utilisé pour les différentes applications du domaine du Traitement Automatique du Langage Naturel (TALN).

Problématique

À la section précédente, nous avons mentionné qu'à cause de la limite du nombre de caractères d'un message imposé par certains médias sociaux, les gens utilisent souvent des façons simples pour exprimer de l'information : un seul mot à la place de plusieurs, une abréviation, un préfixe ou un suffixe, la combinaison de plusieurs mots, etc. En particulier, les caractères chinois permettent d'utiliser des façons simples pour décrire des choses plus complexes. Au chapitre II, nous allons présenter le système d'écriture (les caractères chinois, hanzi) et de représentation de la prononciation (le pinyin). Le processus de conversion des pinyins aux hanzi peut générer diverses erreurs, dont les suivantes :

- Parce que la relation entre le pinyin et le hanzi n'est pas une correspondance un à un, le choix du caractère hanzi correspondant à un pinyin peut varier, ce qui peut causer un mauvais choix de hanzi dans une phrase.
- Parce qu'un pinyin peut correspondre à plusieurs hanzi, les gens peuvent utiliser d'autre hanzi pour remplacer intentionnellement un hanzi dans une phrase. Les autres personnes peuvent néanmoins comprendre la phrase à cause de la prononciation du hanzi, qui est similaire.

- Parce qu'il existe deux systèmes d'écriture pour la langue chinoise, les gens peuvent mélanger les pinyins et les hanzi dans une même phrase. Si le segmenteur ne peut pas identifier les pinyins, il va toujours les traiter comme une séquence de l'anglais, ce qui va en général causer des erreurs.
- Les idiomes chinois possèdent en général quatre hanzi, mais peuvent décrire une longue histoire. Les gens peuvent changer quelques hanzi de l'idiome pour présenter une autre histoire ou une autre chose complexe.

Ces problèmes sont en partie apparus avec les développements plus récents de la langue chinoise, notamment dans les textes de médias sociaux. Le segmenteur de Stanford n'est pas, à l'heure actuelle, adapté pour traiter ces évolutions de la langue. Il s'agit aussi d'une différence entre les textes de journaux et ceux des médias sociaux : dans les textes de journaux, il n'y a pas de mauvais choix de hanzi, ni de remplacement des hanzi dans un mot ou dans un idiome. La segmentation correcte d'une phrase selon son sens original est donc le grand défi de notre thèse. Nous allons maintenant présenter ses défis plus spécifiques.

a. Mélanger les hanzi avec les pinyins ou les symboles

En général, le segmenteur peut faire le prétraitement du corpus, et donc nettoyer les symboles, les alphabets, les signes de ponctuation, etc. Notre système va réaliser ce travail de nettoyage pendant le prétraitement du corpus. Cependant, pour traiter la séquence d'alphabets, notre système ne laissera aucun mot d'anglais ou d'une autre langue, mais plutôt convertir ces séquences vers des hanzi, si possible selon un dictionnaire pinyin vers hanzi. Ensuite, il va segmenter la phrase qui contient les hanzi convertis. Pour le prétraitement des symboles, notre système va les convertir en hanzi selon un dictionnaire symboles vers hanzi, pour ensuite segmenter la phrase qui contient les hanzi convertis.

b. Les mauvais choix de hanzi ou le remplacement des hanzi

Pendant la conversion des pinyins aux hanzi, les gens peuvent faire un choix hâtif, involontaire ou intentionnel. À cause de ces mauvais mots dans une phrase, le segmenteur de Stanford n'arrive souvent pas à segmenter correctement la phrase. Lorsque le texte hanzi contient des erreurs, le système que nous proposons peut convertir les hanzi en pinyins pour trouver les hanzi corrects qui correspondent aux pinyins. Il peut ensuite segmenter correctement la phrase qui contient les bons hanzi. C'est donc dire que notre système peut aussi corriger les fautes d'orthographe.

Contribution

Le travail effectué dans le cadre de cette thèse a pour objectif de proposer un nouveau segmenteur basé sur le segmenteur de Stanford, capable d'améliorer la performance de la segmentation des textes en langue chinoise issus des médias sociaux. Pour cela, nous proposons de combiner la segmentation du pinyin avec la segmentation du hanzi pour améliorer la performance de la segmentation du texte chinois. Les deux segmenteurs font chacun leur travail, selon les règles pour la phrase ; ensuite, le système combine et balance les deux résultats pour obtenir un résultat final. Il s'agit là à notre connaissance de la première utilisation du pinyin pour améliorer la performance de la segmentation des textes écrits en hanzi. Le processus pour convertir les hanzi en pinyins ou les pinyins en hanzi est utilisé pendant la segmentation. Il peut aussi augmenter la performance (vitesse, qualité) de la segmentation.

Le travail a fait trois contributions importantes:

1. La principale (2014, 13 Mai), qui est cognitive, c'est d'avoir reconnu l'importance du pinyin dans la segmentation de la langue chinoise écrite en hanzi notamment dans les microblogs, une forme de la langue écrite plus proche de la langue parlée. Pour segmenter les séquences de hanzi qu'on retrouve dans les textes

écrits en chinois, les systèmes de segmentation actuels se basent sur les hanzi considérés comme des touts (c'est-à-dire sans considérer leurs composantes) et sur la relation entre ces hanzi. Mais les hanzi considérés comme des touts font disparaître la prononciation de la langue et donc les systèmes de segmentation, contrairement aux lecteurs humains du chinois, sont aveugles à tous les aspects de la langue basés sur la phonétique. Dans le mot chinois 拼音 (pīn yīn), il y a le caractère 拼 (pīn) qui signifie « épeler » et le caractère 音 (yīn) qui signifie « son ». La translitération en pinyin, ce n'est donc pas n'importe quelle sorte de translitération mais une translitération qui épèle en sons les mots chinois et donc rend visible aux systèmes de segmentation certains aspects phonétiques auxquels ils étaient jusque-là aveugles. L'étape de translitération en pinyin permet donc à notre système, contrairement à tout autre que nous connaissions, d'être sensible aux dimensions phonétiques du chinois qui sont pertinentes pour segmenter l'écriture.

2. La seconde, c'est l'usage des relations sémantiques entre les mots pour améliorer la segmentation du texte, qui a été implémenté en deux façons : par la structuration du dictionnaire de mots chinois au moyen de la structure sémantique présente dans Haonet (qui est un équivalent chinois de Wordnet) et par l'usage des « semantic chunks ». Grâce à ces ajouts, certaines des relations sémantiques entre les mots qui peuvent influencer la segmentation entre les mots peuvent être pris en considération par notre système lors de la segmentation du texte (et mis en relation avec les facteurs sémantiques).

3. La troisième c'est la pertinence d'utiliser un réseau bayésien pour pondérer, pour chaque mot, la pertinence de chacun de ces facteurs (phonétique, sémantique) mais aussi des autres plus traditionnels (notamment syntaxiques) ou de pertinence locale pour les microblogs (par exemple, les symboles).

La segmentation est un travail complexe. Notre segmenteur considère la probabilité de succession entre les hanzi, la fréquence des mots, la fréquence des hanzi, les erreurs grammaticales, etc. Bien que le segmenteur Stanford considère déjà certaines de ces questions, nous ajoutons le réseau bayésien dans notre système qui a la capacité de prendre en considération simultanément tous les problèmes qui peuvent influencer la performance de la segmentation. À la fin, le modèle proposé du réseau bayésien calculera les probabilités finales pour segmenter la phrase.

Malgré le fait que notre travail soit basé sur le segmenteur de Stanford, celui-ci peut être remplacé par n'importe quel autre système de base. Les composants de notre système peuvent également être adaptés afin d'aménager la possibilité de mettre en place des optimisations futures. En sortie, en plus du résultat de la segmentation du texte, notre système génère un nouveau dictionnaire qui contient de nouveaux mots et de nouvelles expressions rencontrés lors de la segmentation. Ce nouveau dictionnaire peut être réutilisé lors de la prochaine segmentation.

Organisation de la thèse

Cette thèse est organisée en six chapitres :

1. Le premier chapitre, l'introduction, contient la problématique, les objectifs et la contribution de nos travaux pour améliorer la performance de la segmentation.
2. Le second chapitre est une introduction à la langue chinoise, au système d'écriture (le hanzi) et au système de représentation de la prononciation (le pinyin). Avant de commencer la présentation de la segmentation, nous allons présenter les connaissances pertinentes relatives à la langue chinoise.
3. Au troisième chapitre, nous présentons notre première contribution : le segmenteur du pinyin. Dans ce chapitre, nous allons expliquer la conception de la segmentation du hanzi, la conception de la segmentation du pinyin, ainsi que les avantages et les inconvénients des segmenteurs actuels du hanzi lors de la

segmentation de textes de médias sociaux. Les solutions pour résoudre ces problèmes de fond nous mènent à repenser certains concepts de base de la conversion des hanzi aux pinyins et vice-versa. Ces concepts peuvent nous aider à éviter les erreurs pendant la conversion et à améliorer la performance de la segmentation.

4. Au quatrième chapitre, nous faisons état des travaux précédents concernant la segmentation. Ce chapitre rappelle les différents modèles de segmentation du texte chinois. Chaque modèle a des avantages et des inconvénients, que nous discutons.
5. Au cinquième chapitre, nous présentons notre seconde contribution : un modèle hybride de segmentation du hanzi et du pinyin basé sur un réseau bayésien. Ce réseau bayésien considère toutes les influences de la segmentation ; ensuite, il calcule et balance les probabilités entre hanzi et pinyin pour obtenir la meilleure segmentation possible. À cause des différentes influences, ce réseau peut aussi augmenter ou diminuer la probabilité du hanzi dans la phrase.
6. Le sixième chapitre présente et discute les expérimentations que nous avons faites, sur deux différents types de corpus (journaux et médias sociaux) et sur deux tailles différentes chacun, pour valider notre solution.
7. Enfin, le dernier chapitre contient notre conclusion et nos réflexions quant aux perspectives de recherches futures ouvertes par cette recherche.

CHAPITRE I

LA LANGUE CHINOISE

1.1 Introduction

La langue chinoise est une langue parlée par plus de 1,3 milliard de locuteurs. Elle est aussi l'une des six langues reconnues par les Nations Unies.

Notre recherche étant entièrement basée sur la langue chinoise, nous allons en premier lieu présenter, dans ce chapitre, non seulement la langue officielle de la Chine, le mandarin (excluant le cantonais parlé à Hong Kong), mais aussi les autres langues parlées en Chine par différents groupes ethniques, ainsi que les dialectes, comme le dialecte de Shanghai et celui de Wuhan. Ensuite, nous expliquerons l'écriture chinoise et l'histoire de ses caractères. Enfin, nous introduirons le système officiel de la romanisation de la langue chinoise, le pinyin.

1.2 Les langues en Chine

Le *hanyu* (汉语, han4yu3[6], « le chinois ») est une langue qui a une longue histoire. En 2007, le livre 中国的语言[7] (zhong1guo2de0yu3yan2, *La langue chinoise*) explique qu'il existe 129 langues différentes en Chine, sans compter de nombreux dialectes. Il est aussi important de mentionner que la Chine regroupe des ressortissants de 56 nationalités ou groupes ethniques reconnus officiellement. Le *hanyu* est la langue officielle de la République Populaire de Chine et de Singapour. C'est également la langue des Han, le groupe ethnique majoritaire en Chine. En effet, le *hanyu* moderne est la langue utilisée par les Han contemporains. Le *hanyu moderne généralisé* comprend plusieurs dialectes : le *hanyu* est utilisé par les Han, mais aussi, en raison des intégrations culturelles, par d'autres ethnies, qui peuvent le parler dans différentes régions de Chine, avec cependant des différences de prononciation, de vocabulaire et de grammaire. Le *hanyu restrictif*, pour sa part, correspond au *putonghua* (普通话, pu3tong1hua4, « la langue commune »), qu'on appelle aussi le mandarin, lequel est basé sur le dialecte du nord, : sa prononciation standard est celle du dialecte de Pékin et la spécification de sa syntaxe est celle de l'écriture vernaculaire moderne. Enfin la langue officielle du gouvernement chinois est le *putonghua*. On l'appelle aussi communément le *hanyu*, le *zhongwen* (中文, zhong1wen2, « l'écriture chinoise ») ou, à Taiwan, le *guoyu* (国语, guo2yu3, « la langue nationale »).

Le chinois est une langue dite « tonale ». Le ton en chinois réfère à la hauteur (le *pitch*) de la voix lors de la prononciation des syllabes. Le *putonghua* (généralisé et

---

[6] Le mot « han4yu3 » est ici écrit en pinyin. Selon une manière commune d'écrire le pinyin, le chiffre représente le ton de la syllabe. Nous expliquerons un peu plus loin dans ce chapitre le caractère tonal de la langue chinoise ainsi que les différentes façons de représenter les tons en pinyin (chiffres et accents) - voir la page 21.

[7] Ce livre est publié par L'Académie des sciences sociales de Chine (ASSC).

restrictif) possède cinq tons (voir la section 1.2.3 « Le système du pinyin ») : le premier, où la hauteur de la voix est élevée, le deuxième, où la voix est ascendante, le troisième, où la voix descend puis remonte, le quatrième, où la voix tombe rapidement et enfin le ton neutre, où aucune accentuation tonale n'est produite. Chaque syllabe correspond généralement à un caractère sinogramme (ou hanzi (汉字, han4zi4, « le caractère chinois » ; voir ci-dessous), et donc le ton représente la tonalité du caractère.

Une syllabe chinoise dans l'écriture en pinyin, comprend trois parties : la consonne, la voyelle et le ton. La syllabe commence par une consonne, le reste est constitué de voyelles. (Cependant, il faut noter que certaines particules modales ou mots sont dépourvus de consonne, par exemple, 饿 (e4, F : « la faim »), 啊 (a, F : particule modale), etc.) Le ton est alors la tonalité de cette syllabe. Le ton est également considéré comme une partie de la syllabe parce qu'en chinois, les caractères composés des mêmes consonnes et voyelles peuvent être distingués par leur ton. Ainsi, les caractères 汤 (tang1, « la soupe »), 糖 (tang2, « le sucre »), 躺 (tang3, « se coucher ») et 烫 (tang4, « chaud ») » sont composés de la consonne « t » (API : t) et de la voyelle « ang » (composé de la voyelle ɑ et de la consonne nasale vélaire voiléeŋ), mais ces quatre caractères ont des sens différents en raison de leur tonalité différente, ce qui fait d'eux quatre morphèmes différents dans la langue (correspondant à quatre mots différents à l'écrit : 汤, 糖, 躺 et 烫 (« la soupe », « le sucre », « se coucher » et « chaud »).

1.3 L'écriture des langues en Chine

Tel que mentionné précédemment, 129 langues différentes sont parlées en Chine. La plupart d'entre elles utilisent une écriture commune, le hanzi[8] (汉字, han4zi4, « le caractère chinois »), mais les Ouïghour (peuple turcophone) utilisent l'*ouïghour* et les Tibétains utilisent le tibétain, pour ne donner que ces exemples. Ces systèmes d'écriture, l'*ouïghour* et le tibétain, sont très différents du hanzi. Toutefois, puisque le hanzi est le système d'écriture standard pour le *hanyu*, et qu'il est aussi le système d'écriture officiel du gouvernement chinois, dans notre recherche, nous allons simplement aborder le hanzi et son histoire.

Le *hanyu* est un système basé sur le morphème. Les caractères chinois sont unifiés et standardisés. Le *hanyu* moderne aussi possède également une syntaxe unifiée et standardisée. La prononciation du *hanyu* change selon la période de l'histoire, qui est longue, et le territoire, qui est grand : compte tenu de ces facteurs, des dialectes, voire des langues nouvelles, sont apparus. Néanmoins, le fait d'adopter une écriture chinoise commune pour tous ces dialectes et langues élimine les obstacles de communication engendrés par l'apparition de ces nombreuses langues et dialectes différents.

Les caractères chinois ont une forme simplifiée (le caractère simplifié) et une forme traditionnelle (le caractère traditionnel) qu'on appelle 中文字 (zhong1wen2zi4), 中国字 (zhong1guo2zi4) ou 国字 (guo2zi4). Les caractères chinois sont des morphèmes syllabés idéographiques.

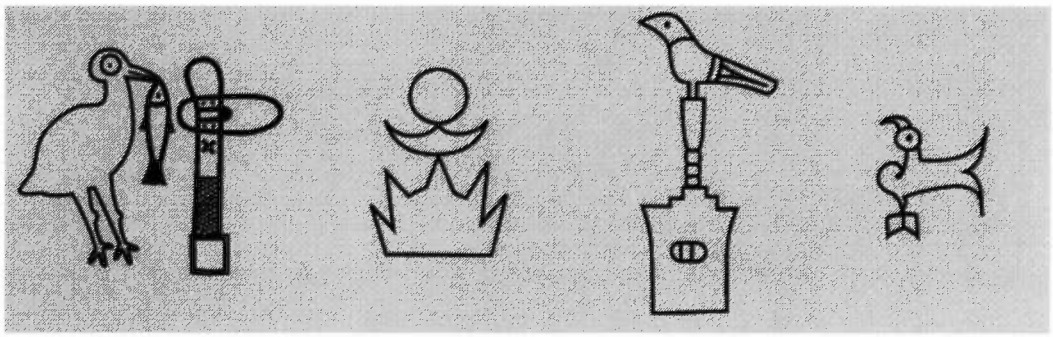
---

[8] Dans ce travail, nous utilisons le mot mandarin 汉字 dans sa forme pinyin simplifiée (sans marque de tons), hanzi, pour désigner les caractères chinois, qu'on appelle également en français « sinogrammes ». Parce que le mot « hanzi » est un emprunt du mandarin, et non un mot français, nous suivons la règle grammaticale du français voulant que les mots empruntés ne reçoivent pas la marque de pluriel du français mais celle de leur langue d'origine. Or puisque le mandarin ne marque pas le pluriel au niveau des mots, nous écrivons « hanzi » au singulier et au pluriel.

Les caractères chinois font partie de l'un des plus anciens systèmes d'écriture. Ils sont utilisés depuis très longtemps dans l'une des plus vastes régions du monde, et ce, par le plus grand nombre de personnes n'ayant jamais utilisé un même système d'écriture.

1.3.1 L'origine des caractères chinois

En 1921, l'archéologue suédois Andersson a trouvé des motifs figuratifs stylisés ou géométriques sur des poteries rouges dans le village de Yangshao de la zone Henan Sanmenxia (Underhill, 2013). Ces poteries rouges ont été fabriquées de 5000 à 3000 AEC (avant l'ère commune). Cette période est identifiée par le nom de la principale culture que l'on retrouvait à cet endroit, Yangshao[9]. Par la suite, les archéologues ont trouvé la culture de Dawenkou (4400 à 2600 AEC), la culture de Liangzhu (3400 à 2250 AEC) et la culture de la dynastie des Shang (1500 à 1000 AEC). La figure 1.1[10] montre un motif utilisé par chacune de ces cultures. De gauche à droite sont représentées la culture de Yangshao, celle de Dawenkou, celle de Liangzhu et la Dynastie des Shang.



1.1 : Les quatre motifs des cultures de chine

[9] https://fr.wikipedia.org/wiki/Culture_de_Yangshao

[10] http://www.ancientscripts.com/chinese.html

Ces motifs peuvent représenter des scènes ou des objets, mais ils ne peuvent pas être considérés comme des caractères chinois. On considère plutôt que ces symboles sculptés apparus au XVI[e] siècle AEC sont à l'origine des caractères chinois. Le système des caractères chinois est un système idéographique, basé sur la pictographie, mais qui utilise aussi des caractères phonétiques. Le nombre total des caractères chinois est d'environ 47 000, dont un noyau d'environ 3 000 caractères plus communs. Avec ces 3 000 caractères, on peut former un très grand nombre des mots, lesquels peuvent constituer à leur tour une variété de phrases.

La construction des caractères chinois suit trois principes :

Principe 1 - *Pictographie*

C'est la première méthode qui forme les caractères primaires. Elle est basée sur une certaine ressemblance entre le caractère et l'objet qu'il désigne. Cette méthode est représentative de la culture chinoise de par son choix de motifs, en utilisant différentes pictographies pour représenter des scènes, sans être un grand motif géométrique. Par exemple :

- 日 (ri4, « soleil » ; originellement ⊙),
- 月 (yue4, « lune » ; originellement),
- 水 (shui3, « eau » ; originellement),
- 牛 (niu2, « bœuf » ; originellement), etc.

Pendant cette période (la culture de Yangshao, 5000 à 3000 AEC), un phénomène intéressant survient, lequel est encore largement présent dans la forme moderne de la langue. Il s'agit de « l'écrit rébus ». Ce phénomène consiste à utiliser un seul caractère pour représenter deux caractères qui ont la même prononciation ou une prononciation similaire, comme les homophones (voir la Figure 2).

1.2 et 1.3 : Deux exemples d'écrit rébus

La pictographie de la figure 1.2 (à gauche) représente les caractères 象 (xiang4, « l'éléphant ») ou 像 (xiang4, « l'image »). Les deux caractères ont la même prononciation, et le deuxième caractère a une partie en commun avec le premier, donc on peut considère que le deuxième est un usage dérivé du premier.

La pictographie de la figure 1.3 (à droite) représente les caractères 王 (wang2, « le roi ») ou 往 (wang3, « vers »). Bien que les deux caractères n'aient pas le même ton, ils ont une prononciation semblable, on peut donc aussi considérer que le deuxième est un usage dérivé du premier.

Après une longue évolution, les pictogrammes ont changé de forme et de structure pour devenir des caractères en position verticale et carrée ; les traits (*strokes*) de certains caractères ont été réduits tandis que d'autres ont été ajoutés. Au terme de cette évolution, les pictogrammes irréguliers sont devenus une police régulière.

Principe 2 - *Idéographie*

Il est facile de créer des pictogrammes, mais ceux-ci ne permettent pas d'exprimer des significations abstraites. Les Anciens ont ainsi créé un autre principe de construction des caractères, l'idéographie. Cette méthode consiste à utiliser différents symboles ou à combiner les pictogrammes avec des symboles dans le but d'exprimer des idées abstraites. Voici un exemple de la première façon de créer un idéogramme, celle qui consiste à créer un symbole à partir d'un pictogramme : 日 (ri4, « soleil » ; originellement ⊙). La deuxième façon de créer un idéogramme, par combinaison de pictogrammes, s'illustre par le cas de l'idéogramme 明 (ming2 ; originellement 𣅀)

qui signifie « lumière » ou « briller », lequel a été créé à partir d'un soleil et d'une lune (qui les deux amènent la lumière). L'idéogramme 旦 (dan4 ; originellement ⊙) signifie « aurore » en montrant un soleil qui se lève à l'horizon. Ces combinaisons permettent d'éviter toute ambigüité de caractères et ainsi permettent de bien distinguer les caractères qui ont la même prononciation ou une prononciation similaire.

Principe 3 - *Idéophonographie*

Il est facile de comprendre la signification des caractères pictographiques et idéographiques, mais on ne peut pas savoir comment les prononcer uniquement à partir de leur apparence ou de ce qu'ils représentent. Ainsi les Anciens ont créé une méthode phonétique pour construire des caractères en combinant un élément phonétique et un élément idéographique. Par exemple, 爸 (ba4, « papa ») est une combinaison de l'élément phonétique 巴 (ba1, « Nul ») et de l'idéogramme 父 (fu4, « père »). Selon les statistiques, les caractères idéophonographiques constituent environ 90 % de l'ensemble des caractères chinois.

Le développement de la civilisation humaine a occasionné le besoin de développer une plus grande variété d'expression et de mots. En utilisant les différentes méthodes mentionnées ci-dessus, les caractères ont ainsi pu donner lieu à une multiplication des expressions linguistiques, et corollairement à la multiplication des significations exprimables par écrit. Par exemple, on utilise le caractère 青 (qing1) seul pour décrire une couleur. Si on veut l'utiliser pour décrire un aspect de l'eau, on utilise 清 (qing1, clair) ; pour décrire un aspect de la température, on écrit 晴 (qing2, « beau temps ») ; pour référer à un insecte, c'est 蜻 (qing1, « libellule ») ; et enfin c'est 请 (qing3, « s'il-vous-plaît ») pour décrire un comportement humain (politesse). La figure ci-dessous présente une comparaison de l'évolution de la forme des caractères chinois à travers le temps.

1.4[11] : L'évolution de la forme des caractères chinois à travers le temps

Selon les circonstances, ces différentes formes de caractères chinois (les polices) sont toujours en usage aujourd'hui. Un site web permet de transformer les caractères en leur forme originale (le cas échéant)[12].

[11] http://www.ancientscripts.com/chinese.html

[12] http://www.diyiziti.com/Builder/114

1.3.2 Les caractères hanzi

L'élément de base de l'écriture chinoise est le trait, c'est-à-dire le processus qui, de la plume baissée à la plume levée, trace un point ou une ligne. Le trait est l'unité minimale pour construire le hanzi. Il existe **six** types principaux de traits : *le trait horizontal* (一), *le trait vertical* (丨), *le trait descendant en s'incurvant vers la gauche* (right-falling) (丿), *le trait descendant en s'incurvant vers la droite* (left-falling) (乀), *le trait tournant* (turning) (亅) et *le point* (丶). Les traits sont tracés suivant un ordre précis et ils sont orientés.

Stroke	Direction	Name	Example
丶	↘	diǎn	你 字 学
一	→	héng	大 三 司
丨	↓	shù	中 上 作
丿	↙	piě	你 千 字
乀	↘	nà	天 人 木
𠂉	↗	tí	海 我 扦
乛	↘	hénggōu	字 爱 家
亅	↓	shùgōu	水 你 景
乚	↘	xiégōu	我 式 忒
𠃌	↓	héngzhé	国 克 票
乙	↓	shùzhé	忙 它 每

1.5[13] :Les traits

Les différents traits se combinent pour construire les différents caractères qu'on appelle hanzi (ce qui signifie « caractère Han »). Les différents hanzi sont à la base de tous les mots chinois. Certains hanzi n'ont pas de signification, tandis que d'autres ont un ou plusieurs sens différents, par exemple :

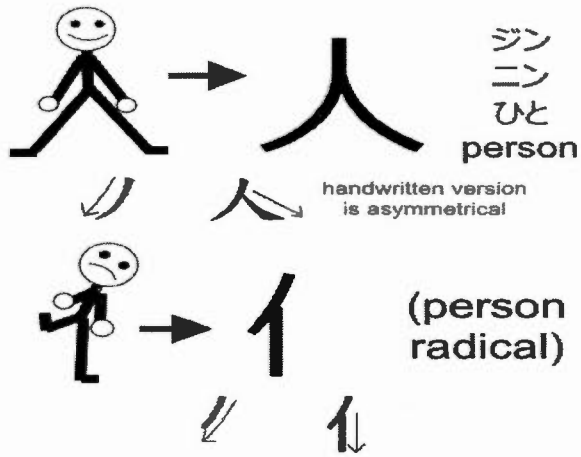
- Le caractère 的 (de5) n'a pas de signification, car il est une particule grammaticale ;
- Le caractère 花 (hua1) signifie « la fleur » ;

[13] « Why Chinese Stroke Order is Important and how to master it ». http://www.digmandarin.com/why-stroke-order-is-important-and-how-to-master-it.html

- Le caractère 重 possède deux prononciations : (chong2) qui signifie « le double » ou « répéter », et la prononciation (zhong4) qui signifie « lourd ».

***Comme dans les autres langues, le caractère chinois est l'unité de base pour représenter un sens ou une idée du monde réel ou abstrait. Le mot chinois peut contenir un seul hanzi (comme le mot 马, ma3, « cheval ») ou plusieurs (comme le mot 中华人民共和国, zhong1hua2ren2min2gong4he2guo2, République Populaire de Chine, qui en contient sept).

On dérive des radicaux de certains caractères très communs. Par exemple, on a ici le mot « personne » (人, ren2) et son radical, tel qu'il apparaît par exemple dans le pronom « tu » (你, ni3) :



1.6[14] : Le hanzi et le radical du 人 (La personne)

La majorité des hanzi est composé par des radicaux, qui peuvent être placés à différents endroits dans le hanzi (à gauche, en haut, ou en bas). En général, un radical a une signification, et le hanzi qui contient ce radical a une signification associée à

[14] http://facets-and-rainbows.tumblr.com/post/91469169095/kanji-人-and-the-person-radical-亻

celle du radical. Par exemple, 足 (zu2, « pieds ») est un radical, et donc la signification du hanzi qui contient ce radical (placé à gauche dans ce hanzi), 跛 (bo3, « boiteux »), est associée à celle du radical. Mais certains radicaux servent comment élément phonétique et non de signification, comme le radical de « cheval » (马) dans « maman » (妈妈).

Bref, le trait est l'unité minimale ; différentes combinaisons de traits construisent le radical ; différentes combinaisons de radicaux ou de radicaux avec d'autres composants construisent le hanzi. En raison du nombre défini de radicaux[15], et aussi parce que le radical constitue un élément de signification important et que plusieurs mots ont des significations reliées, plusieurs hanzi peuvent avoir les mêmes radicaux. Par exemple, 花 (hua1, « fleur »), 草 (cao3, « herbe ») et 菜 (cai4, « légume ») ont le même radical (艹) (placé en haut dans ces hanzi).

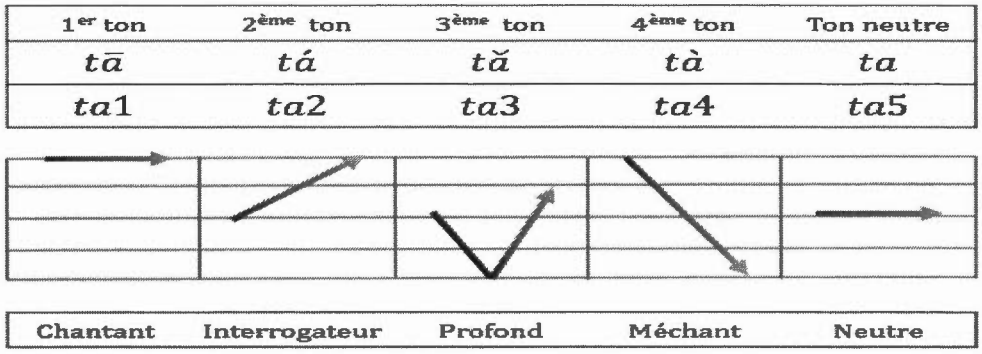
1.3.3 Le système du pinyin

Depuis les années 1950, le pinyin a été développé par le gouvernement chinois pour transcrire les hanzi en caractères de l'alphabet romain. Le système pinyin offre quatre avantages. Premièrement, il est facile à utiliser pour entrer les hanzi dans un ordinateur, dans un téléphone intelligent ou une tablette. Deuxièmement, il permet d'aider les locuteurs dont la langue est basée sur un système alphabétique à apprendre la langue chinoise. Troisièmement, et comme nous l'exposerons plus bas, il est facile à utiliser pour faire une recherche de hanzi dans un dictionnaire. Quatrièmement, le but le pinyin est d'indiquer la prononciation.

Comme nous l'avons expliqué, les langues chinoises sont des langues tonales. Contrairement aux hanzi, le pinyin indique le ton de chaque syllabe. Il y a deux

[15] Il y a 214 radicaux en tout. Pour la liste, voir ici :
http://www.yellowbridge.com/chinese/radicals.php

manières de le faire (voir la Figure 1.6). La première utilise des accents diacritiques : le *macron* (¯) pour représenter le premier ton (où la hauteur de la voix est élevée), l'*accent aigu* (´) pour représenter le deuxième ton (où la voix est ascendante), le *caron* (ˇ) pour représenter le troisième ton (où la voix descend puis remonte), l'accent *grave* (`) pour représenter le quatrième ton (où la voix tombe rapidement) et l'absence *d'accent* pour représenter le ton neutre (où aucune accentuation tonale n'est produite). Par exemple : rén (人, « la personne »), péngyǒu (朋友, « l'ami »), qiǎokèlì (巧克力, « le chocolat ») et Zhōngguó (中国, « la Chine »). Toutefois, pour simplifier l'entrée du pinyin en utilisant un clavier occidental (qu'il soit QWERTY ou AZERTY), on utilise souvent une autre notation, basée sur les chiffres 1, 2, 3, 4 et 5 pour représenter respectivement les cinq tons (¯), (´), (ˇ), (`) et (Null) (voir la Figure 6). La méthode des chiffres étant la plus facile à utiliser pour intégrer le pinyin dans un fichier électronique, comme c'est le cas dans cette thèse de doctorat, c'est la plus souvent utilisée par les Chinois pour écrire les microblogues. Ainsi, selon cette seconde méthode, les mots précédents s'écrivent ren2 (人, « la personne »), peng2you3 (朋友, « l'ami »), qiao3ke4li4 (巧克力, « le chocolat ») et Zhong1guo2 (中国, « la Chine »).



1.7 : Les cinq tons[16]

La marque de ton est une partie indispensable du pinyin, car elle permet de distinguer la signification des mots. Sans marque de ton, il est par exemple impossible de distinguer ces mots différents (et écrits avec des hanzi différents) : 联系 (lian2xi4, « contacter »), 练习 (lian4xi2, « exercice ») et 怜惜 (lian2xi1, « chérir »). Les trois mots sont distingués non pas par les lettres utilisées pour les écrire (dans tous les cas : « lianxi »), mais par le ton de leur prononciation.

On peut résumer ainsi la relation entre le pinyin, les tons et le hanzi : Le pinyin est construit au moyen de l'alphabet romain, additionné d'accents (ou de chiffres) pour représenter les tons. Le pinyin avec le ton nous aide à prononcer la langue chinoise. Une même séquence de lettres peut donc revêtir différents tons. Par exemple la séquence « ma » qui peut être prononcée au troisième ton (pinyin : ma3 ; hanzi : 马, signifiant « cheval ») et qui peut être aussi prononcée au premier ton (pinyin : ma1 ; hanzi : 妈, signifiant « mère »). Un hanzi peut également être associé à une ou plusieurs transcriptions de pinyin, c'est-à-dire qu'il est possible qu'un même hanzi puisse se prononcer de différentes façons (ces hanzi se nomment des « caractères

[16] « Découvrir et comprendre les tons chinois », http://www.un-zeste-de-chine.com/decouvrir-comprendre-tons-chinois/

polyphoniques »). Par exemple, le hanzi 干 peut se prononcer « gan1 » (sec, propre) ou « gan2 » (faire, travailler). Deux transcriptions pinyin qui diffèrent seulement par leurs tons peuvent aussi être associées à différents hanzi. Par exemple « shi4 » (être) et « shi2 » (dix), qui sont associés respectivement aux hanzi 是 et 十. Si différents hanzi ont la même prononciation, ils sont écrits avec le même pinyin (mêmes caractères alphabétiques et mêmes tons), par exemple, les hanzi 握 et 卧, qui sont prononcés de la même manière et sont donc associés au même pinyin « wo4 » mais qui signifient respectivement « tenir » et « être couché ». On voit donc que la relation entre les mots du chinois, les hanzi et le pinyin est bien complexe, et qu'une simple table d'appariement ne suffit pas pour en rendre compte.

En fonction des différentes caractéristiques de la langue chinoise présentées ci-dessus, la recherche des hanzi dans un dictionnaire peut alors se faire selon trois méthodes différentes.

Méthode 1 - *Selon le* pinyin : si on ne sait pas comment écrire un hanzi, mais qu'on sait comment le prononcer, alors on peut le chercher par le pinyin (lequel représente alphabétiquement sa prononciation) dans un dictionnaire selon l'ordre alphabétique. Par exemple, on peut chercher le pinyin « wo3 » (je) sous la lettre « w » dans le dictionnaire.

Méthode 2 - *Selon le trait* : Le principe ici est de calculer le nombre de traits du hanzi que l'on veut chercher. Ensuite, on peut chercher le hanzi dans un dictionnaire selon l'ordre d'écriture et le nombre de traits. Par exemple, le hanzi 我 (wo3, « je ») contient sept traits. On peut le chercher dans la partie 7 du dictionnaire.

Méthode 3 - *Selon le radical* : L'idée est de chercher un hanzi par ses composants dans un dictionnaire. Si un hanzi est très complexe, c'est-à-dire s'il possède plusieurs traits, on peut utiliser cette méthode pour faire la recherche dans un dictionnaire. Tel que préalablement présenté, les hanzi peuvent avoir les mêmes radicaux, et donc cette méthode nous aide à trouver (et à se souvenir) facilement des hanzi. Par exemple, le

hanzi 菜 (cai4, « légume ») est en partie composé du radical ⺾, et donc on peut le chercher dans la partie du radical ⺾ du dictionnaire.

Ces trois méthodes ont chacune leurs avantages propres. Le pinyin peut aider une personne qui connaît la prononciation, mais qui ne sait pas comment écrire le hanzi (elle peut donc utiliser le pinyin pour trouver le hanzi). Le trait et le radical peuvent aider la personne qui lit un hanzi, mais ne sait pas comment le prononcer (elle peut donc utiliser le hanzi pour trouver sa prononciation).

CHAPITRE II

LA SEGMENTATION DES TEXTES CHINOIS

2.1 Introduction

Dans le chapitre précédent, nous avons expliqué la relation entre les hanzi et le pinyin avec le ton. Ces trois éléments (hanzi, pinyin et ton) sont des composants importants du système d'écriture chinois. L'ensemble des hanzi entre deux points constitue une phrase, mais les hanzi dans une phrase sont mis bout à bout sans espace entre eux. Comment alors séparer les hanzi en mots ? C'est là que réside tout le problème de la segmentation d'un texte chinois. De manière similaire, il n'y a pas d'espaces, ni de signes de ponctuation, dans les transcriptions du pinyin en une séquence de lettres alphabétiques. La séparation des lettres alphabétiques composant le pinyin, donc la segmentation du pinyin, est un défi en soi. Dans ce chapitre, nous allons exposer les deux problèmes de la segmentation. Dans la section sur la segmentation du hanzi, nous allons soulever les difficultés souvent rencontrées dans le corpus des journaux, tandis que dans la section sur la segmentation du pinyin, nous aborderons les difficultés soulevées dans le corpus des médias sociaux, incluant les complications apparues récemment. Enfin, dans la dernière partie, nous allons présenter le défi que constituent les symboles de plus en plus présents dans les textes des médias sociaux.

2.2 La segmentation de mots

Comme nous l'avons exposé dans l'introduction, une phrase chinoise n'est pas comme une phrase en anglais ou en français, car il n'existe pas d'espaces entre les mots. Avant de pouvoir comprendre la phrase, les lecteurs doivent alors segmenter la phrase (la séquence des hanzi) en mots à partir de leur connaissance de la langue chinoise. Il est difficile d'entrer un hanzi avec un clavier, car en général, le clavier des ordinateurs, des téléphones intelligents ou des tablettes ne contient que des touches composées de lettres alphabétiques, de chiffres et de certains symboles, ainsi que quelques touches associées à des fonctionnalités. Le clavier ne contient pas de touches pour les hanzi ou les radicaux de hanzi. La conversion entre le pinyin (écrit au moyen de l'alphabet latin additionné d'accents ou de chiffres) et le hanzi pour entrer du texte chinois dans un ordinateur constituait un grand défi de la langue chinoise (Freitas et Meng (2013), Chen et Lee (2000)). L'idée de la conversion du pinyin aux hanzi repose sur la correspondance entre les deux systèmes d'écriture possibles pour les mots du chinois : il s'agit alors d'entrer la version alphabétique du mot (son pinyin) par le clavier, puis de convertir en sa version hanzi la séquence alphabétique entrée, grâce à un système ou à un logiciel auxiliaire. Pour l'utilisateur, l'objectif de ce processus est de choisir le meilleur[17] hanzi correspondant au contexte du pinyin entré. Beaucoup de chercheurs se sont penchés sur cette question. Chen et Lee (2000) ont développé une approche statistique basée sur la phrase pour calculer les probabilités du pinyin en trigramme et ainsi prédire les meilleurs hanzi correspondants. Lucas et Cynthia (2013) ont quant à eux proposé une méthode qui utilise le modèle de Markov caché. Ces deux méthodes de conversion du pinyin au hanzi sont basées sur un corpus pinyin segmenté. Cependant, dans le contexte qui nous intéresse (par exemple, le microblogue chinois entré à l'aide d'un téléphone

[17] Ici, le mot « meilleur » contient deux significations : (1) ce hanzi correspond à une partie de la séquence des lettres alphabétiques entrée ; (2) ce hanzi lui-même ou en combinaison avec le/les hanzi précédent(s) ou suivant(s) a(ont) un sens selon le contexte.

intelligent), les utilisateurs entrent au clavier des caractères alphabétiques, qui correspondent à une séquence de lettres, mais pas à des mots de la langue anglaise ou française (ou toute autre langue alphabétique), car l'entrée se fait sans espaces ni signes de ponctuation, sauf le signe de ponctuation finale. Voici un exemple d'une telle entrée :

woyaochumenlea

Dans ce contexte, le problème de la segmentation du pinyin se pose. Après avoir terminé le processus de conversion du pinyin aux hanzi, les hanzi sont mis bout à bout pour former une phrase. La séquence des hanzi de l'exemple précédent est alors convertie comme suit :

我要出门了啊[18]

Si un utilisateur n'a pas procédé à la conversion du pinyin aux hanzi, il est difficile de comprendre cette séquence de hanzi ; cela soulève le problème de la segmentation de mots.

Certains chercheurs ont effectué des recherches dans ce domaine, parmi lesquels il faut mentionner Xue (2003), Xia (2000), Xue et Shen (2003), Sproat et Shih (1990) et Wu (2010). L'objectif est de segmenter la séquence des caractères du texte chinois en différents mots, parce que le mot est l'unité minimale pour signifier les choses du monde concret ou abstrait, singulier ou général, réel ou imaginaire.

Par exemple :

- Le texte chinois original : AAAAAAAA[19]

[18] Ceci n'est qu'une des nombreuses possibilités de conversion, car il existe beaucoup de choix de hanzi pour un seul pinyin.

[19] Dans ce travail, nous utilisons la lettre majuscule A pour représenter un caractère chinois quelconque.

- Le texte chinois segmenté : AA|AAA|A|AA[20]

Il est à noter que la segmentation sert à séparer les hanzi entre eux au sein d'une séquence de hanzi, mais qu'il ne faut pas séparer les radicaux des hanzi. Ainsi, si un texte chinois est 我们要去电影院 (wo3men2yao4qu4dian4ying3yuan4, « nous allons au cinéma »), nous pouvons segmenter entre les hanzi, mais nous ne pouvons pas faire une séparation *entre les composantes* (le ou les radicaux et les autres parties) *des* hanzi 们 (men2, suffixe qui marque la pluralité), 影 (ying3, « l'image ») et 院 (yuan4, « la cour ») ; par exemple, on ne peut pas séparer au milieu du hanzi 们 (men2, suffixe qui marque la pluralité) en ses deux parties : le radical 亻 (dan1ren2pang2, le radical de la personne) et le radical ou le caractère 门 (men2, « la porte »).

Dans la section suivante, nous allons détailler notre méthode qui combine ces deux types de segmentation pour augmenter la performance de la segmentation dans les textes standard, comme les journaux, et dans les textes des médias sociaux, par exemple les microblogues.

2.3 Problématique

Même après des années de recherche dans le domaine de la segmentation du texte chinois, les différentes méthodes disponibles sont encore assez limitées. Le segmenteur de Stanford[21], par exemple, est souvent utilisé par les chercheurs, mais ses limites sont les suivantes :

– Il intègre un dictionnaire limité, alors que le chinois est en constante évolution. En effet, de nouveaux mots et de nouvelles utilisations de mots apparaissent

[20] Ici est une possibilité de la segmentation

[21] http://nlp.stanford.edu/software/segmenter.shtml

constamment. Or le dictionnaire du segmenteur de Stanford ne peut pas faire une mise à jour à temps.

- Le dictionnaire du segmenteur de Stanford est un dictionnaire général, c'est-à-dire qu'il contient beaucoup de mots dans différents domaines, or ces mots sont le plus souvent utilisés par domaine, et le segmenteur ne spécifie par le ou les domaines relatifs à chaque mot.
- En chinois, il existe beaucoup d'idiomes[22] qui contiennent quatre caractères. Par exemple, le segmenteur de Stanford segmente toujours les idiomes au milieu, deux par deux. C'est un inconvénient du segmenteur de Stanford. D'autres segmenteurs traitent ce problème d'idiomes.

Les nouveaux mots, les nouvelles expressions ou les nouveaux idiomes apparaissent dans les textes selon l'époque et selon le médium (article, film, série de télévision, etc.). Ces dernières années, Facebook et le « microblogging » (comme Twitter, Facebook ou Weibo en Chine[23]) deviennent aussi de plus en plus populaires. Toute personne peut les utiliser pour publier une information, faire un commentaire ou communiquer avec d'autres personnes. En raison des limites spatiales (par exemple, les fameux 144 caractères de Twitter) imposées aux utilisateurs, ces différents écrits contiennent souvent de nouveaux mots. Ces plateformes sont aussi un grand défi pour la segmentation du texte chinois non seulement à cause de la création constante de mots nouveaux, mais aussi parce que le langage y est utilisé de manière beaucoup plus familière. Après une analyse des travaux déjà publiés sur le sujet, nous résumerons les cinq problèmes les plus communs de la segmentation du texte chinois. Ensuite, nous proposerons notre approche, qui est basée sur les outils existants pour améliorer la performance de la segmentation du texte chinois.

---

[22] La Définition de l'idiome, voir « *Le problème des nouveaux idiomes* » ci-dessus.

[23] http://d.weibo.com

1. Le problème des nouveaux mots ou des nouvelles expressions

Le problème de la création constante de nouveaux mots et de nouvelles expressions a déjà été étudié par Tanabe, et al. (2014). En voici les deux aspects principaux :

- Le premier aspect dépend des nouvelles combinaisons apparues entre les différents hanzi. Les nouveaux mots de ce type sont pour la plupart des noms propres. Par exemple, le mot 土豪金 (Tu3Hao2Jin1) signifiant littéralement « l'or des propriétaires ruraux » (*rural landlords gold*) a été créé pour décrire la nouvelle couleur or de l'iPhone 5S. L'origine de la construction de ce mot est l'idée que l'or est une couleur qui symbolise la noblesse, la richesse et la prospérité. L'apparition de ce mot nouveau accompagne celle de choses nouvelles. Avec le temps, certains de ces mots nouveaux sont de moins en moins utilisés et disparaissent, mais le mot 土豪金 (Tu3Hao2Jin1) continue à être utilisé (pour l'Iphone 6, ainsi que pour d'autres outils technologiques dispendieux).
- Le deuxième aspect consiste à utiliser un ou plusieurs hanzi pour remplacer un autre ou plusieurs autres hanzi dans un mot déjà existant. Par exemple, le mot 十面埋伏 (Shi2Mian4Mai2Fu2) réfère à la situation folklorique d'une troupe embusquée et encerclée par des ennemis, mais qui réussit néanmoins à les anéantir. Le mot est aussi devenu le titre d'un film, *Le Secret des poignards volants* (*House of Flying Daggers*). Le nouveau mot 十面霾伏 (Shi2Mian4Mai2Fu2) a la même prononciation que le mot précédent, et il est donc signifié par le même pinyin. Il remplace simplement le hanzi 埋 (Mai2) par 霾 (Mai2). L'idée derrière ce mot nouveau est la suivante : à cause de la pollution qui sévit partout en Chine, beaucoup de villes sont encerclées par un nouvel ennemi, le smog. Dans la langue chinoise, le mot 雾霾 (Wu4Mai2) décrit la brume sèche ; on a donc utilisé la composante 霾 (Mai2)

de ce mot pour remplacer la composante 埋 (Mai2) du premier mot. Le nouveau mot réfère ainsi à la brume sèche polluée (le smog) encerclant les différentes villes de la Chine.

2. *Le problème des nouveaux idiomes*

Un idiome est une locution figée qu'on utilise habituellement depuis longtemps. Les idiomes proviennent par exemple des classiques anciens (par exemple les *Analectes* de Confucius), des écrits historiques ou des histoires orales. La signification des idiomes est profonde et elle est souvent implicite par rapport au sens littéral des termes. De plus, cette signification est plus que la somme des termes constituants l'idiome, elle est holistique. La structure de l'idiome est compacte (réduite au hanzi minimal pour exprimer le sens) et solide (on ne peut pas ajouter, supprimer ou remplacer des hanzi). En chinois, la plupart des idiomes contiennent quatre hanzi. Selon les besoins des gens, de nouveaux idiomes sont créés en combinant différents hanzi, souvent en groupes de quatre pour suivre la tradition. Par exemple, le nouvel idiome 不明觉厉 (Bu4Ming2Jue2Li4) signifie la situation psychologique complexe « bien que je ne comprenais pas ce qu'il voulait dire, je sentais néanmoins que son propos était redoutable ».

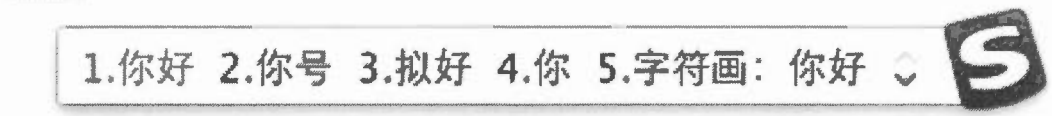
3. *Le problème des abréviations*

Comme la langue française, qui utilise par exemple « s.v.p. » comme abréviation de « s'il vous plaît », la langue chinoise utilise aussi des abréviations, lesquelles peuvent causer des problèmes pour la segmentation. Par exemple, le mot 中非 (zhong1fei1) a deux significations selon le contexte : il peut signifier la *République centrafricaine* (le pays), ou encore il peut signifier de manière abrégée les relations sino-africaines (中国-非洲, zhong1guo2- fei1zhou1). Cette seconde

expression est formée à partir du premier hanzi du nom populaire de la Chine (中国- , zhong1guo2-) et du premier du nom de l'Afrique (非洲, fei1zhou1).

4. Le problème du pinyin

Tel qu'expliqué plus haut, le pinyin est une façon d'entrer les hanzi dans un système informatique (ordinateur, téléphone intelligent, tablette). Il existe plusieurs outils qui permettent de convertir du pinyin aux hanzi. Toutefois, pour le moment, nous allons uniquement évoquer l'idée qui se trouve derrière ces outils. Ensuite, nous identifierons les problèmes qui peuvent survenir au cours du processus de conversion. La figure 7 ci-dessous illustre l'un de ces outils[24].

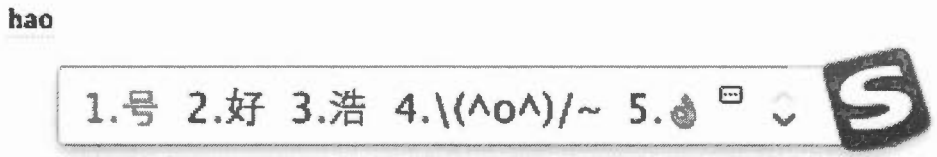


2.1 : Un outil de conversion

La figure 2.1 montre que l'utilisateur entre, par le biais de son clavier, le pinyin avec les lettres alphabétiques, et que le système propose alors un choix de hanzi correspondant à l'idée qu'il veut écrire.

[24] http://pinyin.sogou.com/mac/introduce.php

2.2 : Les hanzi correspondant au pinyin « ni »



2.3 : Les hanzi correspondant au pinyin « hao »

Le processus de conversion suit les étapes suivantes :

1. Une personne entre par le biais de son clavier les lettres alphabétiques du pinyin en séquence.
2. L'outil segmente cette séquence de lettres alphabétiques en mots pinyin au moyen d'un symbole particulier, comme le symbole « ' » entre les lettres « i » et « h » dans l'exemple précédent. De cette façon, l'outil propose uniquement l'ensemble des lettres « ni » pour former un pinyin, et la même chose pour « hao ».
3. Le pinyin de « ni » et de « hao » correspond chacun à plusieurs choix de hanzi, comme indiqué dans les figures 2.2 et 2.3. Dans ces exemples, l'outil a été configuré pour afficher cinq choix par page (à la fin de la ligne, on voit la flèche pour la page suivante). Parmi les choix suggérés, le plus fréquent est mis en avant.

4. L'outil propose ainsi des choix de mots pour chacune des deux transcriptions en pinyin de « ni » et de « hao ». Lorsque la personne choisit un numéro, le processus de conversion est terminé.

2.4 La segmentation du pinyin

Comme nous l'avons mentionné dans la section de la segmentation du hanzi, la conversion du pinyin aux hanzi est aujourd'hui un processus très important dans la langue chinoise, étant donné que ce processus est déterminant pour la compréhension d'un texte. L'idée de la segmentation du pinyin est premièrement de segmenter la séquence de caractères inscrits en lettres alphabétiques en différentes unités, chaque unité correspondant à un mot pinyin, sans qu'il reste un ou plusieurs caractères en lettres alphabétiques qui ne correspondraient à aucun pinyin.

Nous utilisons la segmentation du pinyin dans notre projet parce que nous considérons que cette segmentation contribue à la performance de la conversion du pinyin aux hanzi. Cette segmentation permet de proposer des choix de hanzi inexacts pendant la conversion. Par exemple, si pendant la conversion, l'utilisateur choisit un hanzi inexact par mégarde, il va ensuite le remplacer par le bon hanzi qui correspond au sens qu'il veut donner. Mais si ensuite le segmenteur segmente ces hanzi en mots, le résultat de la segmentation comportera une erreur. Nous jugeons que cette segmentation est un facteur important qui peut influer sur la performance de la segmentation du hanzi dans les textes des médias sociaux.

Si on reprend notre exemple précédent :

- Étant donné la séquence des lettres : woyaochumenlea (je vais sortir)

La mauvaise segmentation est w|o|yao|chu|men|l|e|a[25], où les caractères de lettres alphabétiques « w » et « l » ne correspondent à aucun pinyin, même si les caractères de lettres alphabétiques « o », « e » et « a » correspondent de fait à un pinyin.

Deuxièmement, si nous ne tenons pas compte du contexte ou du sens de la phrase lors de la segmentation du pinyin, une séquence de lettres alphabétiques possède plusieurs résultats segmentés différents. Les différentes unités segmentées correspondent aux hanzi. Si les hanzi peuvent construire les différents mots qui contribuent au sens du contexte, alors cette segmentation sera jugée satisfaisante. Si ce n'est pas le cas, il faudra alors ajuster la position de la segmentation entre les lettres alphabétiques.

Par exemple:

- Il existe deux façons de segmenter la séquence des lettres alphabétiques « woyaochumenlea » (je vais sortir) avec la possibilité de construire le pinyin comme suit :

  1. wo|ya|o|chu|men|le|a : Dans cette séquence segmentée, le sens des hanzi qui correspondent au pinyin « ya » et « o » n'apporte aucune contribution au sens de la phrase ou au contexte.

  2. wo|yao|chu|men|le|a : Dans cette séquence segmentée, les hanzi qui correspondent au pinyin « yao » et « chu » apportent une contribution au sens de la phrase et au contexte (leur signification est « vais sortir ») et donc cette segmentation sera jugée acceptable.

L'exemple précédent concerne une phrase, l'exemple suivant concerne maintenant un mot.

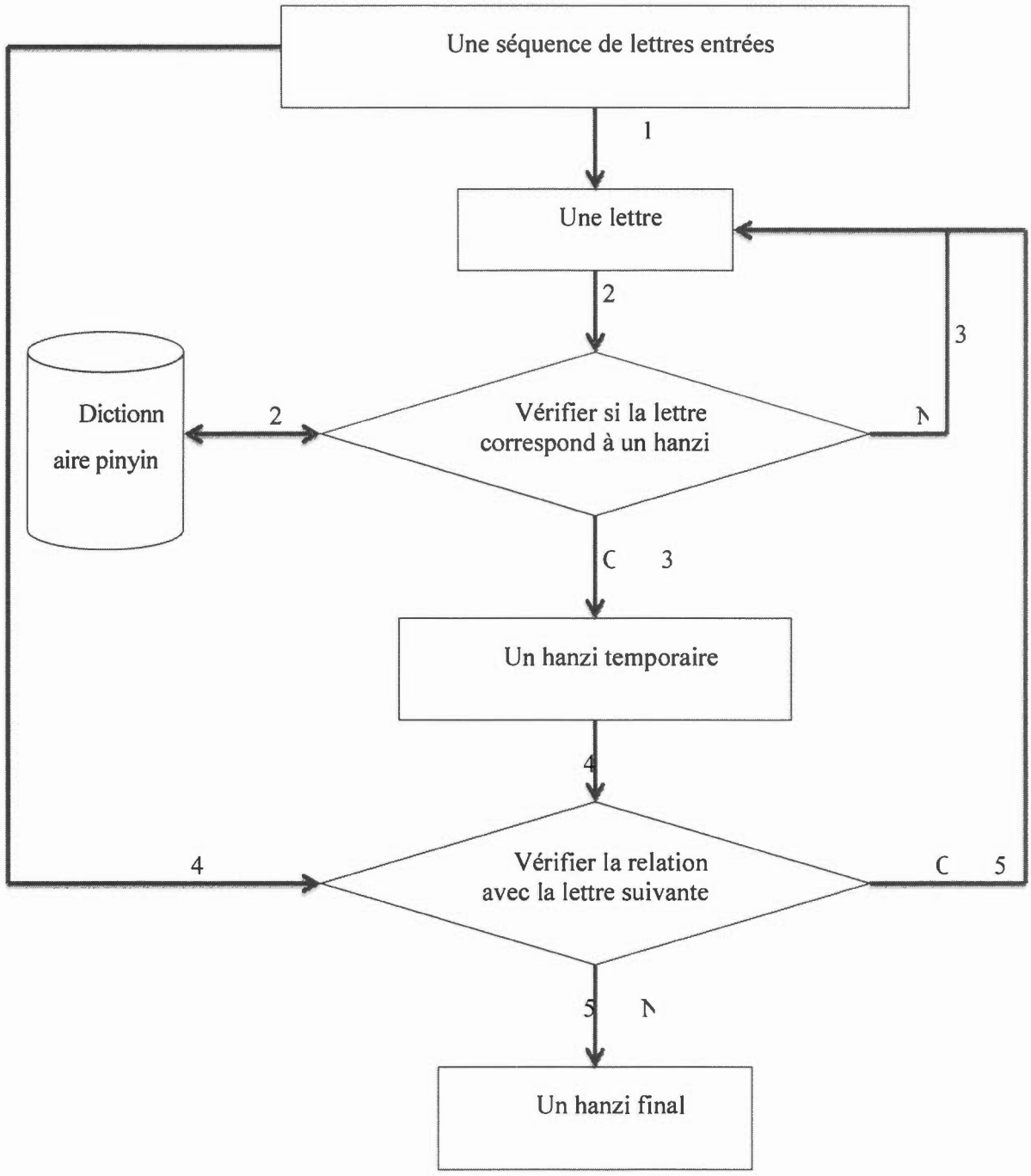
Par exemple:

[25] Dans ce travail, nous utilisons la barre verticale « | » pour indiquer l'endroit d'une segmentation.

- Il existe deux façons de segmenter la séquence de lettres « wanan ». La première est « wan|an » et l'autre est « wa|nan ». Ces deux résultats correspondent à deux façons différentes de choisir des hanzi correspondant au pinyin. Segmentées de la première façon, soit « wan|an », l'unité correspondante à un mot en pinyin signifie « bonne nuit » ; or en revanche, la seconde segmentation de la séquence ne produit pas une unité ayant un sens : l'ensemble du pinyin n'a alors aucun sens.

Tel qu'expliqué au chapitre II, le pinyin est un système qui permet de romaniser le caractère hanzi avec des lettres alphabétiques, tout en en conservant le sens. Le pinyin et le hanzi auront le même sens puisqu'ils se correspondent exactement. Comme expliqué précédemment, le pinyin est une façon de représenter alphabétiquement les hanzi. Si un pinyin segmenté a un sens, alors le hanzi correspondant a le même sens.

Nous examinerons ici la façon de convertir une séquence de lettres alphabétiques en un hanzi (voir la figure 2.4). Dans ce processus, une séquence de lettres alphabétiques est entrée par le moyen d'un clavier (processus 1 dans la figure). Le système de la conversion compare ces entrées avec celles d'un dictionnaire du pinyin (processus 2 dans la figure). Si l'ensemble des lettres alphabétiques entrées correspond à un hanzi, le processus garde en mémoire le hanzi correspondant (processus 3 dans la figure). Dans le cas contraire, les lettres alphabétiques entrées sont associées avec la lettre suivante pour vérifier à nouveau dans le dictionnaire du pinyin pour compléter le processus de conversion. Si ce hanzi (du processus 3) associé avec la lettre suivante ne correspond pas à un nouveau hanzi, alors le premier choix de hanzi va être conservé (processus 4 et 5 dans la figure) ; dans le cas contraire, le système de conversion retourne au processus 2.



2.4 : Le processus de conversion de pinyin à hanzi par clavier

Au cours de ce processus, trois types de problèmes, qui auront un impact subséquent sur la segmentation, pourraient survenir :

1. Lors du processus de conversion, l'outil propose des choix, mais l'utilisateur fait un choix hâtif (ce que les gens font souvent pour gagner du temps lors de l'entrée du texte) ou tout simplement une erreur de frappe. C'est ce qu'on appelle aussi le problème de la polyphonie[26]. En conséquence, l'utilisateur entre un autre hanzi, ou même un autre mot, à la place du véritable mot qu'il voulait écrire. Cette erreur va influer sur la performance de la segmentation du texte chinois en créant des séquences de hanzi qui ne se retrouvent pas dans les dictionnaires ou en affectant le calcul de probabilités entre les hanzi.
2. Une deuxième situation pourrait se produire si l'utilisateur choisit intentionnellement un hanzi ou un mot différent pour remplacer le mot qu'il veut écrire : c'est ce que plusieurs utilisateurs font pour entrer du texte chinois dans le système par l'ordinateur ou la tablette afin de jouer avec les mots. Dans ce type de cas, la performance de la segmentation va être affectée pour les mêmes raisons que celles énumérées ci-dessus, bien que l'origine des mauvais choix de hanzi soit différente. Beaucoup de travaux ont été menés sur les erreurs et les choix hâtifs, parmi lesquels il faut mentionner Li et Peng (2011), Yang, Zhao *et al.* (2012), Jia, Wang et Zhao (2013), Liu *et al.* (2013), Zheng, Li et Sun (2011), et Zheng, Xie *et al.* (2011), mais peu de travaux se sont penchés sur le cas de la modification intentionnelle.
3. Un troisième cas que nous retrouvons souvent dans les microblogues ou sur Weibo est apparu très récemment et prend la forme suivante :

   Texte chinois : 各种让你看了都爆 (hai) 炸 (pa)的瞬间

[26] Certains hanzi peuvent être prononcés de plusieurs façons : on les dits polyphoniques (多音字, duo1yin1zi4). Par exemple : le hanzi 乐, qui peut se prononcer le4 (content) et yue4 (la musique), est dit polyphonique.

Ce texte, qui contient à la fois des hanzi et du pinyin, offre deux possibilités de lecture. Soit on ne lit que les hanzi et on obtient le sens suivant :

- Les différents moments d'explosion pendant que vous lisez[27].

Soit on remplace les hanzi qui précèdent directement par le pinyin pour obtenir alors le sens suivant :

- Les différents moments de peur pendant que vous lisez.

Cette nouvelle façon d'écrire consiste à utiliser un pinyin derrière un hanzi et à répéter cette opération à deux reprises. L'ensemble des hanzi possède alors une signification, et il en est de même pour la phrase qui résulte lors de l'utilisation du pinyin pour remplacer les hanzi uniques qui les précèdent. Cette forme de composition va bien entendu engendrer des complications lors de la segmentation parce qu'elle sépare des hanzi normalement pris dans leur ensemble et aussi parce qu'il y a deux façons distinctes d'analyser la syntaxe du texte écrit.

2.5 Le problème des symboles

Enfin, la cinquième difficulté soulevée lors du processus de segmentation concerne la présence de chiffres et de symboles dans les textes. Nous avons vu que l'outil nous propose un choix de hanzi ou de mots. Dans l'exemple de la figure 2, nous voyons cependant que le quatrième choix est un symbole. La situation est la même dans l'exemple de la figure 3 : le quatrième et le cinquième choix sont des symboles. Ces derniers vont engendrer des complications lors de la segmentation, car ils se substituent souvent à des mots, mais ne se retrouvent pas dans les dictionnaires.

Le deuxième aspect de ce problème concerne les chiffres. Les utilisateurs de microblogues et de Weibo emploient beaucoup de chiffres, dont la prononciation

[27] Ce sens est évidemment métaphorique, mais on peut imaginer ce que l'auteur a voulu dire.

ressemble à celle des hanzi. Par exemple, puisque la prononciation du chiffre 8 en mandarin est semblable à celle du mot « bye » en anglais, on utilise souvent 88 (ba1ba1) pour représenter 拜拜 (bai1bai1, « au revoir » (F[28]), « bye-bye » (A)).

Un autre exemple est le chiffre 520 (wu2er4ling2), qui est utilisé pour représenter 我爱你 (wo3ai4ni3, « je t'aime ») parce que la prononciation du premier ressemble un peu à celle du second. Ces types d'utilisation de chiffres employés dans un texte vont compliquer la segmentation du texte, d'une part parce qu'une partie du sens de la phrase n'est pas exprimée par des mots, et d'autre part parce que cette utilisation des chiffres renvoie à une autre signification que leur usage habituel (comme si en français on disait « Il fait beau, 520 et la vie est belle »).

Ces cinq types de difficultés identifiées apparaissent souvent dans les textes des médias sociaux et elles ont un impact important sur la performance de la segmentation. Le chapitre IV propose un modèle hybride utilisant les réseaux bayésiens pour résoudre ces différents problèmes.

[28] Ici, nous utilisons F pour indiquer le sens français et A pour indiquer le sens anglais.

CHAPITRE III

ÉTAT DE L'ART

3.1 Introduction

La langue naturelle est la méthode la plus élémentaire, la plus directe et la plus naturelle pour exprimer les pensées et les sentiments humains. Elle est l'outil de communication le plus utilisé dans la société humaine. Un problème important de l'étude du langage est la question de savoir comment modéliser de manière computationnelle la relation entre le langage naturel et les connaissances et objectifs des locuteurs. La recherche en Traitement Automatique du Langage Naturel (TALN) est devenue récemment un outil clé pour comprendre la communication linguistique entre les personnes, et entre les personnes et les machines. Cette recherche multidisciplinaire, qui a rallié plusieurs chercheurs au cours des dernières années, fait principalement appel aux sciences cognitives, à l'informatique, à la linguistique et aux mathématiques.

La première étape du traitement de la langue chinoise est la segmentation du texte. Elle en est même une composante nécessaire dans plusieurs autres applications en TALN, par exemple la recherche d'informations (*Information Retrival*), le résumé automatique de texte (*Automatic Text Summarization*), la correction automatique d'épreuves (*Automatic Proofreading*), l'entrée intelligente des hanzi, la conversion

entre le chinois simplifié et le chinois traditionnel, la traduction automatique, la synthèse vocale, etc.

Dans ce chapitre, nous allons d'abord présenter le premier système de segmentation CDWS (The Modern Printed Chinese Distinguishing Word System) et son schéma. Nous allons aussi présenter certains articles traitant des recherches initiales dans ce domaine. Ensuite, nous classerons et présenterons les méthodes de segmentation actuelles selon leur algorithme. Ces algorithmes peuvent être classés en quatre catégories :

1) la segmentation mécanique;

2) la segmentation basée sur les règles;

3) la segmentation statistique;

4) la segmentation hybride.

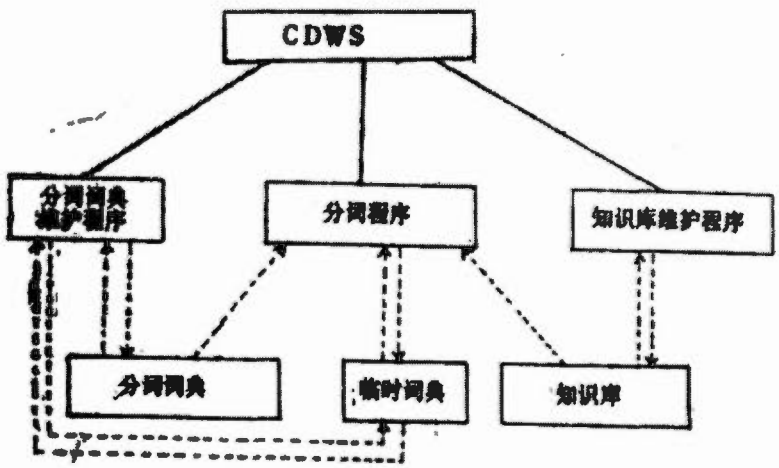
Finalement, nous présenterons les travaux antérieurs portant sur la segmentation du texte des médias sociaux. Les limites et les inconvénients de ces travaux seront discutés en conclusion.

3.2 Le premier système de la segmentation CDWS

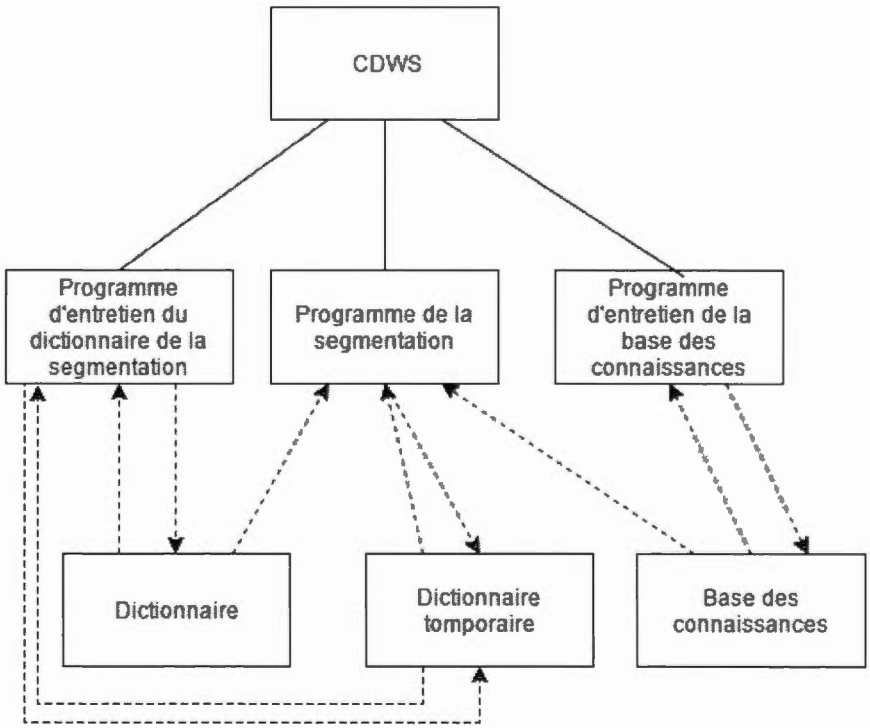
C'est Liang (1987) qui a proposé le premier système de segmentation, le *Modern Printed Chinese Distinguishing Word System (CDWS)* qui a été développé au début des années 1980 par l'université Beihang, également appelée Université d'aéronautique et d'astronautique de Pékin ou BUAA. Le segmenteur CDWS utilisait la méthode du *Forward Maximum Matching (FMM)*[29], un algorithme de détection des erreurs à la position suffixe d'un mot et de correction des connaissances. Le dictionnaire intégré du CDWS contient 124 500 mots, parmi lesquels des noms

[29] Nous présentons cette méthode à la section suivante.

propres, des abréviations, des idiomes, des expressions, des structures POS (par exemple : V. + objet) et certains OOV. Le CDWS propose un modèle computationnel du hanzi, comme le démontrent la figure 3.1 (modèle en chinois) et la figure 3.2 (modèle en français). Ling (1987) a aussi proposé les premières définitions des concepts de segmentation automatique du texte chinois. Ce système est le premier travail de catégorisation des hanzi et des mots ambigus.



3.1 : Le premier système de segmentation CDMS



3.2 : Le premier système de la segmentation CDWS

Le premier article scientifique portant sur la segmentation du texte chinois est apparu en 1990. Effectivement, Sproat et Shih (1990) ont proposé une approche statistique pour détecter la frontière des mots. Cette approche regroupe le texte chinois par groupes de deux caractères en calculant la probabilité mutuelle de chaque paire. Puisqu'elle considère ainsi que chaque mot chinois contient seulement deux caractères, elle ne peut pas traiter les mots qui en contiennent plus de deux, comme les idiomes, les noms propres, les expressions, etc.

Sun et Xu (2011) ont pour leur part employé des algorithmes qui utilisent un dictionnaire et des règles pour segmenter le texte chinois.

3.3 La segmentation mécanique

L'idée à la base de la segmentation mécanique consiste à apparier (*matching*) la séquence des hanzi en utilisant le dictionnaire. Selon la direction de ce *matching*, le processus de segmentation peut être divisé en *Forward Matching*, *Reverse Matching* et *Bi-directional Matching*. Selon l'importance que l'on accorde aux mots plus longs ou plus courts, la segmentation peut être divisée en *Maximun Matching* et *Minimum Matching*. Nous présentons dans les sections qui suivent les différentes méthodes qui combinent ces formes d'appariement.

3.3.1 La méthode Forward Maximum Matching (FMM)

La stratégie derrière cette méthode consiste à utiliser le nombre maximal de hanzi formant au moins un mot du dictionnaire pour segmenter une première séquence de hanzi, comme l'ont fait Wang, Deng et Zou (2006). Si l'ensemble de hanzi présentement segmenté ne correspond pas à un mot dans le dictionnaire, la méthode recommence en utilisant le nombre maximum de hanzi moins 1 pour refaire la segmentation et la comparaison avec le dictionnaire. Pendant ce processus, si un mot est trouvé, ce mot est retiré de la séquence à segmenter et la méthode recommence avec le nombre maximum sur les hanzi suivant le mot trouvé. Si le nombre devient 0 et qu'aucun mot n'est trouvé, le système classe le premier caractère de la séquence comme OOV dans le dictionnaire et recommence ensuite le processus avec le nombre maximum à partir du hanzi suivant.

Si par exemple, nous avons la séquence de hanzi « 1234567 »[30] et que, dans le dictionnaire utilisé, les mots contenant le plus de hanzi en contiennent 5, alors le

[30] Ici, chaque chiffre présente un hanzi. Cette séquence contient 7 hanzi.

processus se déroule comme décrit ci-dessus jusqu'à ce que la segmentation se termine :

Étape 1 : Commencer au début et voir que la séquence 12345 n'est pas un mot.

Étape2 : Réduire le nombre maximal de hanzi considérés de 1 à 4. Si la séquence 1234 correspond à un mot, le processus va continuer avec 567 et revenir à l'étape 1. Si la séquence 1234 n'est pas un mot, alors le processus refait l'étape 2, et ce, jusqu'à ce que le nombre maximal devienne 0.

Les articles suivants utilisent ou sont basés sur cette méthode, qui est fondamentale pour la segmentation du texte chinois. Bien que le système CDWS de l'université Beihang ait proposé la maquette de cette méthode, à ce moment-là, elle était encore incomplète.

Chen et Liu (1992) ont proposé un algorithme de *matching* utilisant six règles heuristiques pour identifier les mots. Leur approche pose cependant plusieurs problèmes. Premièrement, elle est basée sur la langue cantonaise et elle n'utilise pas le même système de pinyin que celui qui est officiellement reconnu en Chine aujourd'hui, ce qui n'empêche pas l'emploi des règles heuristiques pour le mandarin. Deuxièmement, cette méthode considère que la segmentation la plus plausible des mots est de trois caractères, et ne peut donc pas résoudre les mots qui sont construits par deux sous-mots. Ainsi, cette méthode considère que la meilleure segmentation de l'idiome de quatre caractères est deux fois deux caractères.

Liu et al. (1994) ont proposé la version définitive de cette méthode. Dans leur article, ces auteurs démontrent que la précision de la segmentation peut atteindre jusqu'à 95 % si le dictionnaire est plus large et plus complet. Un peu plus tard, Wong et Chan (1996) ont proposé un algorithme basé sur le FMM et la force des liens entre les caractères ou les mots. Ils ont analysé la force entre le caractère individuel (qui est aussi un mot) et les caractères préfixes et suffixes dans un mot (soit les premiers et derniers caractères du mot). Si ce caractère individuel peut construire un nouveau mot

avec le préfixe ou le suffixe, la segmentation en entier sera modifiée. David et Marti (1997) ont fait une expérimentation pour la recherche d'informations en utilisant trois différents algorithmes, dont le FMM. Ensuite, ils ont comparé leurs résultats et évalué leurs performances lors de la recherche d'informations ; c'est le FMM a obtenu la deuxième place. Tang, Wu et Li (2015) ont proposé un nouvel algorithme basé sur le FMM, lequel construit un nouveau mécanisme du dictionnaire. À l'opposé de la méthode FMM, le système, à chaque itération, ajoute un hanzi et utilise une fonction de désambiguïsation pour effectuer la segmentation.

3.3.2 La méthode Reverse Maximum Matching (RMM)

Le *Reverse Maximum Matching (RMM)* est obtenu en renversant la direction du *Maximum Matching (MM)*. Ainsi, cette méthode commence à la fin de chaque phrase. Ding, Zhang et Li (2009) ont proposé un nouveau fonctionnement du dictionnaire en utilisant une structure de la table de hachage pour améliorer le RMM. En particulier, cette nouvelle méthode a amélioré la vitesse et la précision de la segmentation. Mo et al. (2013) ont à leur tour développé un nouveau fonctionnement du dictionnaire qui utilise une double table de hachage pour améliorer les vitesses du RMM et du FMM, ce qui a permis d'augmenter la performance de la segmentation.

En 2014, Liu et Li (2014) ont proposé un algorithme de *Reverse Backtracking* basé sur le RMM. Cette méthode utilise un algorithme pour déterminer en premier le mot présentant un nombre maximum de hanzi dans la séquence qui constitue la phrase. Si ce n'est pas le cas, elle enlève le premier hanzi de la séquence et recommence le processus. La segmentation se déroule selon l'ordre du nombre des hanzi du mot. Si cette méthode a permis d'augmenter la vitesse de traitement, elle rencontre des problèmes avec une séquence de mots comme AAABBB. Ainsi, si le mot AABBB comporte 5 caractères, il va le traiter en premier et le résultat de la segmentation sera

A|AABBB, alors que la segmentation exacte aurait dû contenir deux mots : AAA et BBB.

En 2012, Jiao et Peng (2012) ont travaillé avec un corpus tiré des médias sociaux. Ils ont proposé une méthode pour déterminer le thème principal d'un texte de microblogue. La méthode utilise le RMM pour capturer les informations importantes, pour ensuite les classifier et analyser les opinions publiques.

3.3.3 La méthode Bi-directional matching

La méthode *Bi-directional Matching,* combinant le résultat du FMM et du RMM, choisit la meilleure segmentation en la comparant avec le dictionnaire, et donne ainsi un résultat bien meilleur à la segmentation, comme l'ont révélé les travaux de Wang, Qian, Zhao, Song, He et Wu (2016). Selon les recherches de Sun et Benjamin (1995), la combinaison du FMM et du RFM donne un résultat absolument identique et satisfaisant pour 90 % des phrases. Pour 9 % d'entre elles, un résultat sur deux est probant en utilisant la fonction de désambiguïsation. En revanche, des travaux supplémentaires sont nécessaires quant au 1 % des phrases pour lequel le résultat n'est pas satisfaisant.

Au regard des travaux de désambiguïsation effectués en 1989, Wang, Wang, Li et Bai ont proposé la méthode *Minimum Sub-Lexical*, qui cherche à trouver une façon de segmenter le texte contenant un nombre minimal de mots. Les autres travaux importants sur cette question ont été proposés en 1996 par Ma, avec la méthode *Full Syncopation*. La méthode consiste à énumérer toutes les segmentations possibles, ce qui permet d'éviter les ambiguïtés tout en créant, malheureusement, plusieurs segmentations inutiles.

3.4 La segmentation basée sur les règles

L'idée de base des travaux de Zhang et al. (2012), de Ma et Hinrichs (2015) et de Mark et Katherine (2010) repose sur l'utilisation d'une l'analyse sémantique et syntaxique pour traiter l'ambiguïté.

En 1997, Chang et Su (1997) ont proposé une méthode non supervisée itérative en utilisant les contraintes contextuelles et une matrice de caractères conjoints pour extraire les nouveaux mots (OOV[31]). Cette méthode permet la mise à jour des nouveaux mots dans un dictionnaire, et l'utilisation ultérieure de ce dictionnaire lors d'un nouveau processus de segmentation. Ce processus est basé sur un dictionnaire supplémentaire qui permet d'identifier les mots.

3.5 La segmentation basée sur les statistiques

Dans le corpus, lorsque deux hanzi sont adjacents, il est plus probable qu'ils constituent un mot. Cette méthode permet de calculer les probabilités entre les hanzi adjacents, et si la probabilité est plus proche de 1, cela signifie que deux hanzi adjacents forment un même mot. En revanche, si la probabilité est plus proche de 0, il est peu probable alors que les deux hanzi adjacents forment un mot (sauf peut-être un nouveau mot inconnu). Les chercheurs Gao, Wang, Li et Lee (2000), Zhang et al. (2016), Liu et al. (2016), Cai et Zhao (2016), Sharon et al. (2006), Mochihashi et al. (2009), Chen et al. (2014), Magistry (2012), Xu et al. (2008), Zhao et Kit (2008) et Wang et al. (2011) ont développé une technique permettant de calculer la probabilité entre les hanzi, entre un hanzi avec un mot, ou encore, entre un mot avec un autre mot.

---

[31] *Out Of Vocabulary (OOV)*. Il s'agit des mots qui apparaissent dans le test, mais qui ne se retrouvent pas dans les bases de donnée et recueils de vocabulaires usuels, les dictionnaires par exemple.

Sun, Shen et Tsou (1998) ont proposé une méthode qui calcule le t-score de la probabilité des caractères en bi-gramme, tri-gramme et 4-gramme. Cette méthode est désormais un module pleinement intégré dans plusieurs outils de segmentation.

Zhang, Yu, Xiong et Liu (2003) ont proposé une technique basée sur le modèle hiérarchique de Markov caché (*Hierarchical Hidden Markov Model, HHMM*). Le HHMM est basé sur le modèle de Markov caché (*Hidden Markov Model, HMM*) utilisant un quintuple pour étiqueter et segmenter la séquence des hanzi. Ces deux techniques, HHMM et HMM, sont basées sur la méthode de Xia (2000) décrite ci-dessous.

Liu, D., et al. (2009) ont proposé une méthode basée sur le réseau bayésien en alignant les caractères avec l'algorithme Viterbi. Cette méthode considère le premier caractère d'un mot comme un état, et la probabilité n-gramme est comprise comme la probabilité de transition a priori entre les éléments du n-gramme (la probabilité a priori du troisième caractère étant donné les deux premiers, du quatrième étant donné les trois premiers, etc.) Cette méthode peut bien identifier le nom propre, par exemple, le nom de la personne ou le nom d'une entreprise, etc., mais est très limitée quand elle utilise la probabilité bigramme pour identifier la relation entre deux caractères ou un caractère avec un mot.

3.6 Segmentation hybride et autres

Nie et al. (1995) ont créé une méthode hybride combinant à la fois la méthode basée sur un dictionnaire avec des règles de grammaire et la méthode statistique. Selon leur article, les deux méthodes fonctionnent indépendamment. La méthode statistique est utilisée sur le résultat de la méthode, basée elle-même sur le dictionnaire pour déterminer les OOV. Si les OOV n'apparaissent pas souvent dans le

corpus, la méthode n'est alors pas en mesure de déterminer ces OOVs, comme le démontre les travaux de Lu (2007).

En 2000, Xia a révélé le *Penn Chinese Treebank,* qui a été utilisé dans le segmenteur de Stanford jusqu'à aujourd'hui. Levy et Manning (2003) ont également étudié cette technique. Le *Penn Chinese Treebank* est un corpus segmenté et POS *Tagged*[32] (selon les travaux de Zheng et al. (2013)), qui contient environ 5 000 000 mots chinois. En 2001, Peng et Schurmans ont proposé le modèle des champs aléatoires conditionnels *(Condition Random Fields* ou *CRF).* Zhong et al. (2012) et Zhao et al. (2006) ont également effectué des recherches dans ce même domaine. Le système CRF est dynamique, car il évalue différents facteurs, comme la fréquence des mots et le contexte. Jusqu'à présent, plusieurs outils de la segmentation sont basés sur ce modèle. Cependant, ce n'est qu'en 2002 que la segmentation du texte chinois est passée à un nouveau niveau. Effectivement, Xue et Converse (2002) ont utilisé des étiquettes (*tags*) pour labelliser les caractères, en les répertoriant sous quatre catégories. Ces étiquettes se définissent par « L » pour *left*, « R » pour *right* et « M » pour *middle*. Ainsi un caractère est identifié par l'une ou l'autre de ces quatre étiquettes (LL, RR, MM, LR) selon les règles suivantes :

- LL, si la frontière d'un mot est à gauche de ce caractère, et si ce caractère peut construire un mot avec un (des) caractère(s) à droite ;
- RR, si la frontière d'un mot est à droite de ce caractère, et si ce caractère peut construire un mot avec un (des) caractère(s) à gauche ;
- MM, si ce caractère est au milieu d'un mot et
- LR, si ce caractère peut construire un mot à lui tout seul.

Xue et Shen (2003) et Xue (2003) ont développé davantage cette technique de labellisation en proposant quatre nouvelles étiquettes pour les caractères :

[32]Une étiquette *Part-Of-Speech* (POS) associe des mots du texte à des informations grammaticales correspondantes (comme le rôle syntaxique, le genre ou le nombre, etc.), lesquelles contribueront à l'analyse grammaticale et syntaxique. Voir ci-dessous la description du module de POS-*tagging* de notre système.

- LR, si ce caractère apparait à la frontière à droite d'un mot ;
- LM, si ce caractère apparait à la frontière à gauche d'un mot ;
- MM, si ce caractère apparait au milieu d'un mot et
- MR, si ce caractère construit un mot à lui tout seul (LMR-*tagging*).

Cette approche utilise les règles générales pour identifier les positions des caractères, sans toutefois les appliquer aux nouveaux mots ou aux expressions qui n'ont pas encore été labellisés.

Chen, Qiu et Huang (2017) (2015) ont trouvé que les modèles récents basés sur les réseaux neuronaux ne peuvent pas extraire les compositions complexes dans une phrase avec les caractéristiques discrètes. Ils ont proposé un modèle neuronal qui enrichit la fonction conjointe de la segmentation et du POS tagging. En particulier, ce modèle utilise les informations portant sur les dépendances à longue distance pour modéliser les caractéristiques de composition complexe dans une phrase. Les expérimentations sur cinq différents corpus ont montré l'efficacité de ce modèle.

Le tableau suivant compare la performance de trois types de méthodes de segmentation : celles qui se basent sur un dictionnaire, celles qui se basent sur les règles et enfin celles qui se basent sur la statistique.

Le tableau 3.3 montre que la segmentation basée sur un dictionnaire n'est pas efficace pour identifier un OOV, c'est-à-dire que si un mot n'est pas dans le dictionnaire, cette méthode va considérer qu'il est un OOV. Si le dictionnaire n'est pas mis à jour fréquemment, alors le résultat de la segmentation contiendra plusieurs OOV. La segmentation basée sur les règles peut identifier les OOV au moyen par exemple de la vérification grammaticale (mais aussi d'autres règles pertinentes). La segmentation statistique peut elle aussi bien identifier les OOV en fonction des probabilités entre les mots dans un corpus.

Les difficultés présentées par les trois méthodes sont donc différentes. À cause de la variété de la langue, il est difficile de constituer une base pour la segmentation

basée sur les règles. À cause de la taille des corpus utilisés pour entraîner les segmenteurs, le temps de l'entraînement et la quantité de probabilités qu'il faut enregistrer dans la base de données sont deux grands défis pour la segmentation statistique. Enfin, la segmentation basée sur un dictionnaire est facile à réaliser, et le temps pour effectuer la segmentation est plus court que celle des deux autres méthodes, mais cette méthode n'est pas efficace pour les OOV.

Méthodes	Basé sur un dictionnaire	Basé sur les règles	Statistique
Désambigüisation	Faible	Fort	Fort
OOV	Faible	Moyenne	Fort
Dictionnaire	Nécessaire	Pas nécessaire	Pas nécessaire
Corpus	Pas nécessaire	Pas nécessaire	Nécessaire
Base de règles	Pas nécessaire	Nécessaire	Pas nécessaire
Complexité	Facile	Difficile	Moyenne
Maturité	Oui	Non	Oui
Difficulté	Facile	Difficile	Moyenne
Précision	Moyenne	Précise	Précise
Vitesse	Rapide	Lente	Moyenne

3.3 : La comparaison entre les méthodes

La technique de la segmentation statistique utilise au début du processus l'algorithme sans dictionnaire. Il s'agit donc d'une segmentation sans dictionnaire,

utilisant uniquement les probabilités. Cette technique permet la mise à jour d'un dictionnaire répertoriant les mots segmentés qu'il a préalablement identifiés. Ainsi, ces mots déjà segmentés peuvent être réutilisés lors d'une prochaine segmentation.

3.7 Les travaux sur les corpus des médias sociaux

Ces dernières années, il y a eu beaucoup de travaux dans le domaine de la segmentation, comme ceux de He, He, Cen et Lu (2012), Liu et al. (2012), Qiu et al. (2015). Wang et Yu (2010), Wang, Yu, Zhu Et Li (2012), Dang et Valette (2017). Ces chercheurs ont construit une base de données contenant environ 6 790 idiomes et permettant leur identification automatique dans le texte chinois. Li (2011) a proposé une approche combinant à la fois l'analyse lexicale et l'analyse syntaxique. Cette approche enrichit le *Penn Chinese Treebank POS tags*. En ce qui concerne le pinyin, Chen et Lee (2000), ainsi que Lucas et Cynthia (2013), ont travaillé sur la conversion du pinyin au hanzi. Ils ont essayé de traiter les ambiguïtés pendant le processus de conversion, comme les polyphonies et les homophonies. De leur côté, Yang et al. (2012), Ying et al. (2012) et Li et Peng (2011) se sont attachés à la vérification de l'orthographe. Leurs travaux sont basés sur les règles grammaticales. Leurs techniques respectives permettent de détecter et corriger les erreurs des hanzi.

Trois chercheurs se sont penchés plus particulièrement sur le domaine des médias sociaux. Zhang et al. (2014) ont développé une approche permettant de détecter la présence d'un sujet d'actualité dans les médias sociaux en calculant la fréquence des mots. Song et al. (2012) ainsi que Wei (2009) ont quant à eux mis en place une segmentation du texte tiré de médias sociaux utilisant la probabilité. Enfin, Leonardo (2004) a essayé de construire une relation sémantique entre les mots dans une même phrase, une approche considère la relation entre les phrases : si un mot est identifié dans la phrase précédente, il pourra influer sur la probabilité que ce mot se retrouve dans la phrase suivante. Yuan et Purver (2012) ont proposé une méthode pour

déterminer les émotions contenues dans les textes des médias sociaux. Les méthodes existantes ne permettant pas de distinguer les émotions du contexte, ces chercheurs ont classé des émotions et déterminé des mots susceptibles d'en exprimer.

3.8 Conclusion

Comme nous l'avons démontré au chapitre précédent, les textes des médias sociaux en chinois contiennent beaucoup de pinyins, d'émoticônes et de nouveaux mots. Si les méthodes existantes arrivent à bien segmenter les textes normalisés, comme ceux des journaux par exemple, lorsqu'on utilise ces méthodes pour des textes tirés des médias sociaux, la qualité de la segmentation diminue de beaucoup. Ces méthodes n'arrivent pas à détecter les mauvaises segmentations ou les mauvaises utilisations de hanzi.

CHAPITRE IV

MODÉLISATION BASÉE SUR LE RÉSEAU BAYÉSIEN

4.1 Introduction

Dans ce chapitre, nous allons présenter notre système de segmentation dans son intégralité et effectuer une comparaison entre les différents systèmes. Ce système contient plusieurs composants, tels que HowNet par Dong (2010), le réseau bayésien, le segmenteur du pinyin et le segmenteur de Stanford. Chaque composant réalise sa propre fonction au sein du système complet. Nous n'allons présenter ici que les composants que nous avons développés ou adaptés.

D'abord, nous allons présenter le dictionnaire sémantique basé sur le HowNet par Dong (2010). Il existe aussi beaucoup de travaux basés sur HowNet, comme ceux de Li et al. (2009), Dai et al. (2008), Huang et al. (2008), Wang (2002) et Zhang et al. (2012).

La plupart des travaux utilisent HowNet pour calculer la similarité des mots ou des phrases, ou pour analyser les tendances des sentiments dans les textes chinois. Notre système combine Hownet avec un dictionnaire supplémentaire33 pour construire un

[33] Ici nous utilisons un dictionnaire qui contient 2 million de mot comme un dictionnaire supplémentaire du site web NLPIR : www.nlpir.org.

dictionnaire sémantiquement structuré pour notre système (lequel nous permettra d'utiliser des relations sémantiques entre les mots pour améliorer la segmentation du texte chinois). Nous allons calculer la similarité des mots du dictionnaire supplémentaire en utilisant les concepts et les sémèmes d'HowNet. Le but de ce calcul est de regrouper les éléments du dictionnaire en groupes de mots contenant des mots similaires. Dans notre projet, chaque groupe de mots contient aussi les mots qui ont le même caractère avec la même prononciation. Les détails du regroupement sont expliqués à la section 5.2.

Ensuite, nous allons présenter le cœur de notre système : le réseau bayésien. Ce réseau combine les différentes influences du côté de la segmentation du hanzi et du côté de la segmentation du pinyin pour évaluer la probabilité de satisfaction d'une segmentation. Ce réseau continue son travail en boucle jusqu'à ce que cette probabilité se stabilise. La segmentation du hanzi segmente le texte en mots. La segmentation du pinyin segmente la séquence alphabétique en pinyins. Dans les textes des médias sociaux, il existe beaucoup de hanzi incorrects (voir la section 3.4). Selon le résultat de la segmentation du pinyin et en le comparant avec le résultat du hanzi, nous pouvons trouver les hanzi corrects par rapport au dictionnaire du pinyin-hanzi. Cette étape peut aider à améliorer la segmentation du hanzi. Ici, nous ne nous sommes pas simplement contenté d'utiliser les hanzi corrects (trouvés par la segmentation du pinyin) pour remplacer les hanzi incorrects et refaire la segmentation par le segmenteur de Stanford. Nous utilisons le réseau bayésien en considérant les différents facteurs des deux côtés (hanzi et pinyin), pour calculer une satisfaction de la segmentation qui convient aux deux côtés (voir ci-dessous la section 5.5).

À la fin de ce chapitre, nous présenterons enfin notre système dans son ensemble. Au prochain chapitre, nous comparerons la performance de notre système à celle de trois autres systèmes de segmentation pour deux types de corpus : le domaine journalistique et celui des médias sociaux.

4.2 HowNet

Dong (1999) a construit une base de connaissances appelée Hownet, laquelle contient une variété de relations entre les concepts chinois et leurs attributs. Ce chercheur décrit Hownet ainsi : *"HowNet is a fully computational knowledge base providing computer-readable knowledge that is crucial to text understanding and machine translation."* Plus récemment, Dong et al. (2006) ont défini le HowNet comme *"[...] an on-line common-sense knowledge base unveiling inter-conceptual relations and inter-attribute relations of concepts as connoting in lexicons of Chinese and their English equivalents"*. Le réseau HowNet de Dong et al. (2003) est basé sur l'ensemble de sémèmes représentant les unités fondées sur le sens indissociable. Chaque concept d'HowNet est défini par des sémèmes ou d'autres concepts. La structure d'HowNet prend la forme d'un graphe, et non d'une hiérarchie ou d'une arborescence. HowNet contient un graphe décrivant des relations inter-conceptuelles et des relations inter-attributs. L'ensemble des nœuds (les concepts ou les attributs) représentés par le graphe permet de démontrer les propriétés générales et spécifiques des concepts. Voici un exemple tiré de HowNet:

Le concept : 地铁 (di4tie3, le métro)

La définition (DEF) :

1. def : {Landvehicle|车 :
   {transport|运送 :
   instrument = {~},
   location = {location|位置 :
   belong = {land|陆地 :
   modifier = {beneath|下}}},
   patient = {human|人}}}

2. def : {Facilities|设施 :

location = {location|位置 :

belong = {land|陆地 :

modifier = {beneath|下}}},

{transport|运送 :

instrument = {~},

patient = {human|人}}}

Dans cet exemple, le mot 地铁 (di4tie3, le métro) est défini de deux façons. Les unités de définitions sont des sémèmes, comme « transport|运送 (yu4song4)[34] », « land|陆地 (lu4si4) », et « human|人 (ren2) », etc. Dans les définitions, il y a aussi des relations, comme *"location"*, *"belong"* et *"instrument"* pour représenter les relations sémantiques entre les concepts.

- Dans la définition 1, le sémème « Landvehicle|车 (che1) » est l'attribut catégoriel; il est aussi l'hyperonyme du concept « 地铁 (di4tie3, le métro) ». Le sémème « transport|运送 (yun4song4) » est l'hyperonyme du concept « Landvehicle|车 (che1) ». Les sémèmes « ~[35] », « location|位置 (wei4zhi4) » et « human|人 (ren2) » sont les attributs spécifiques du concept « Landvehicle|车 (che1) ». Ces trois attributs amènent les informations supplémentaires du concept. Les sémèmes « land|陆地 (lu4di4) » et « beneath|下 (xia4) » sont les attributs spécifiques du concept « location|位置 (wei4zhi4) ». Ils donnent les informations plus précises du concept « location|位置 (wei4zhi4) ».

---

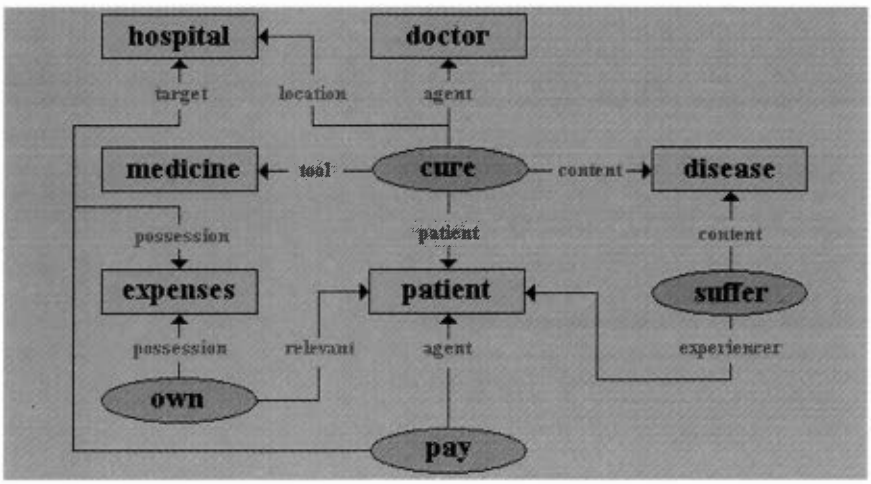
34 Ici, nous ajoutons le pinyin pour chaque concept, n'existant pas le système Hownet.

35 ~ représente le concept du début : 地铁 (di4tie3, le métro).

- Dans la définition 2, le sémème « Facilities| 设 施 (she4shi1) » est l'hyperonyme du concept « 地铁 (di4tie3, le métro) ». Le sémème « location| 位 置 (wei4zhi4) » est l'attribut local du concept « Facilities| 设 施 (she4shi1) ». Les sémèmes « land|陆地 (lu4di4) » et « beneath|下 (xia4) » sont les attributs spécifiques du concept « location|位置 (wei4zhi4) ». le sémème « transport|运送 (yun4song4) » est l'attribut spécifique du concept « Facilities|设施 (she4shi1) ». Les attributs « ~ » et « human|人 (ren2) » sont les attributs spécifiques du concept « transport|运送 (yun4song4) ».

Dans cet exemple, le sémème « transport|运送 (yun4song4) » a deux différentes fonctions, l'une comme l'hyperonyme d'un concept dans la définition 1, l'autre, comme un attribut spécifique d'un concept dans la définition 2 ; ceci étant lié aux choix proposés parmi les différentes relations pour définir le concept « 地铁 (di4tie3, le métro) ». Dans les deux définitions, les différents sémèmes représentent leurs différentes relations avec les autres concepts ou les attributs.

Nous avons reproduit ici une figure illustrant les relations entre le patient, le médecin et l'hôpital.



4.1 : L'exemple du graphe du HowNet

À la figure 4.1, chaque rectangle vert représente une entité, chaque ovale violet représente un événement, et enfin, chaque flèche représente une relation. La relation local-événement entre *" hospital "* et *"cure"* signifie que la personne est guérie à l'hôpital. La relation agent-événement entre *"doctor"* et *"cure"* précise que le médecin a effectué l'action de « soigner », et la relation patient-événement entre *"patient"* et *"cure"* indique que le patient a accepté l'action « soigner ».

Selon les graphes, nous voyons que les relations sont le pivot d'HowNet, qui est constitué de deux parties : 1) l'ensemble des sémèmes de base (l'entité, la partie ou le composant, l'attribut, la valeur d'attribut, l'événement, le temps et l'espace); 2) les relations entre les sémèmes listées ci-dessous :

- superordinate-subordinate (上级 (shang4ji2) -下级 (xia4ji2), supérieur-subordonné)[36]
- synonym (同义词 (tong2yi4ci2), synonyme)
- antonym (反义词 (fan4yi4ci2), antonyme)
- converse (逆的 (ni4de), converse)
- part-whole (部分-整体 (bu4fen1-zheng3ti3), partie-entier)
- attribute-host (属性-主体 (shu3xin4-zhu3ti3), attribut-hôte)
- material-product (材料-产品 (cai2liao4-chan3pin3), matériel-product)
- agent-event (施事-事件 (shi1shi4-shi4jian4), agent-événement)
- patient-event (受事-事件 (shou4shi4-shi4jian4), patient-événement)
- instrument-event (设施-事件 (she4shi1-shi4jian4), instrument-événement)
- location-event (地点-事件 (di4dian3-shi4jian4), local-événement)
- time-event (时间-事件 (shi2jain1-shi4jian4), temps-événement)
- value-attribute (值-属性 (zhi2-shu3xing4), valeur-attribut)

[36] Ici, nous avons choisi d'ajouter le hanzi, le pinyin et le français.

- entity-value (实物-值 (shi2wu4-zhi2), entité-valeur)
- event-role (事件-角色 (shi4jian4-jue2se4), événement-rôle)
- concepts related (概念相关 (gai4nian4xiang1guan1), concepts liés)

Les tableaux 4.2 et 4.3 qui suivent présentent les caractéristiques d'HowNet [37].

HowNet	Chinois	Anglais
Caractère	20 892	-
Mot et Expression	122 078	111 945
Signification	138 957	132 369
Définition	32 137	32 137

4.2 : La taille du HowNet

POS tagging	Chinois	Anglais
Adj.	13 721	12 599
Adv.	2 433	3 122
Aux.	103	102
Caractère	14 405	0
Classifier	447	0

[37] https://groups.google.com/forum/#!forum/hownet

Conj.	135	85
Coor.	14	7
Det.	58	123
Echo	137	7
Expression	850	1000
Infs.	0	7
Lettre	57	57
Nom	61 712	66 238
Numéral	557	572
pp	0	1372
Préfix	7	28
Prep.	257	320
Pron.	176	109
Pun.	45	50
Root	0	3 233
Stru.	82	0
Suffixe	0	7
Verbe	32 620	27 415
Local	48	87
Total	116 540	127 864

4.3 : Nombre des catégories syntaxiques

HowNet	Chinois	Anglais
Entité	3	4
Événement	15558	14799
Attribut	5201	4665
Valeur de l'attribut	10967	10860
Chose	94593	94579
Temps	2872	2872
Espace	1427	1427
Composant	9209	9208
Total	138414	139830

4.4 : Nombre des catégories sémantiques

Ces chiffres du tableau 4.4 révèlent qu'HowNet a déjà déterminé plus de 130 000 concepts. Cependant, au regard du développement de la langue chinoise, HowNet a aussi dû se développer pour tenir compte des nouveaux mots et des nouvelles expressions, ce qui a été fait depuis les travaux de Gan et Wong (2000).

Comme nous l'avons expliqué auparavant, le graphe d'HowNet connecte les différents sémèmes associés, c'est-à-dire ceux qui sont liés par des relations. Nous considérons donc que cet ensemble de sémèmes a des relations sémantiques. Au chapitre II, nous avons expliqué que certains caractères chinois ont des significations

alors que d'autres n'en ont pas. Ainsi, si un mot contient un ou certains caractères qu'un autre mot possède aussi, alors il est possible que les deux mots possèdent un ou plusieurs sens communs implicites, un ou plusieurs sens dérivés (voir l'exemple ci-dessous). Considérant cette caractéristique des sinogrammes, nous allons utiliser le graphe d'HowNet pour nous aider à améliorer la performance de la segmentation du texte chinois en nous basant sur ces relations sémantiques entre les mots.

Exemple :

Le caractère 的 a quatre différentes prononciations :

- « de3 » : il existe cinq utilisations.
  1) Utilisé après un mot ou une expression pour représenter un adjectif, par exemple : 美丽的 (mei3li4de3, F : beau).
  2) Utilisé pour référer une chose ou une personne comme un pronom, par exemple : 唱歌的 (chang4ge1de3, F : ce chanteur).
  3) Utilisé pour exprimer la relation appartenue, par exemple : 他的衣服 (ta1de3yi1fu2, F : son vêtement).
  4) Utilisé, en particulier, à la fin de la phrase comme une particule. Il est souvent utilisé avec le caractère « 是 » (shi4, F : oui) pour représenter un ton positif, par exemple : 你是对的 (ni3shi4dui4de3, F : Tu es correct).
  5) Utilisé après un mot ou une expression pour représenter un adverbe. Il est utilisé comme le mot « 地 » (di3, F : I, comme un adverbe, 2, le sol).
- « di1 » : Utilisé pour représenter le taxi, et les choses ou les personnes associées au taxi. Par exemple : « 的士 » (di1shi4, F : le taxi), « 的士司机 » (di1shi4si1ji1, F : le chauffeur du taxi), etc.
- « di2 » : Utilisé pour représenter « vraiment », par exemple : 的确 (di2que4, F : vraiment, c'est vrai).

- « di4 » : Utilisé pour exprimer le centre de pare-balles, par exemple : 有的放矢 (you3di4fang4shi1, F : tirer la flèche sur la cible, c'est-à-dire, avoir un objet défini en vue), et 目的 (mu4di4, F : l'objectif à atteindre), etc.

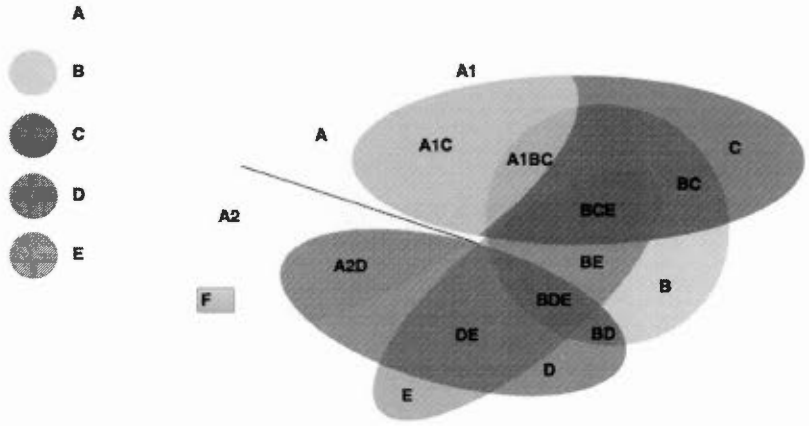
Dans cet exemple, nous voyons que certains caractères possèdent différentes prononciations, correspondant à différentes utilisations et différents sens.

Par exemple, si le caractère « 的 » est utilisé comme un adjectif pour représenter une couleur, donc notre système de la segmentation en se combinant au HowNet va les mettre dans un même groupe. Si le caractère « 的 » se prononce « di1 », le système va les mettre dans un même groupe. Ces deux groupes sont liés par le caractère « 的 ».

HowNet va y contribuer pour améliorer la performance de la segmentation sur deux aspects, en l'utilisant pour :

i. Regrouper le dictionnaire utilisé lors de la segmentation.
ii. Mettre en œuvre le module *Semantic Chunk*.

Pour ce projet, nous utilisons un dictionnaire supplémentaire qui contient plus de deux millions de mots chinois. Évidemment, si le système doit calculer les probabilités pour chacun de ces mots, il va prendre beaucoup de temps. Cependant, si nous regroupons préalablement ces mots tel qu'illustré à la figure 4.5, le temps de calcul diminuera, optimisant ainsi l'efficacité du système.



4.5 : Regroupement du dictionnaire en utilisant le système HowNet

Voici les étapes du regroupement du dictionnaire :

a) L'ensemble des mots possédant un même caractère est regroupé. La figure 4.5 montre, par exemple, que le groupe A en jaune décrit tout le groupe de mots contenant le même caractère A, alors qu'AC représente le mot avec les deux mêmes caractères AC.

b) Cet ensemble de mots est divisé selon le pinyin du caractère. Ainsi A1 et A2 ont le même caractère (A, par exemple 长), mais ils ont un pinyin différent (par exemple, chang1 (longueur, long) et zhang3 (grandir, augmenter)). Comme nous l'avons expliqué au chapitre I, la langue chinoise est très polyphonique et un même caractère peut avoir une ou plusieurs transcriptions en pinyin. Ces différentes transcriptions représentent différentes significations du caractère. L'ensemble de mots A2D possède le même caractère A, le même le pinyin A2 et le même caractère D.

c) L'ensemble des mots du graphe d'HowNet. Dans le système HowNet, un mot a des relations sémantiques avec les autres mots. La figure 4.5 ci-dessus

montre qu'un mot qui contient A se retrouve dans le même regroupement que le mot F ayant la même sémantique.

Après ces trois étapes, le grand dictionnaire est regroupé en petits ensembles, ce qui peut aider le module mot clé et le module Structure-Phrase (les détails vont être présentés pendant la présentation des modules).

Pendant le processus de regroupement du dictionnaire, il existe des mots qui vont apparaître dans plusieurs groupes à cause de chacun de leurs caractères. Mais à cause de la taille du HowNet, ça ne créé pas trop de groupes et n'influence pas trop négativement la vitesse de la segmentation. En revanche, cette étape peut être utilisée pour le module Semantic Chunk en utilisant la structure du HowNet pour créer un dictionnaire sémantique (voir l'explication ci-dessus). Le résultat de la segmentation du pinyin peut parcourir ce dictionnaire pour trouver les hanzi corrects.

Ensuite, le module *Semantic Chunk* représente une deuxième application possible d'HowNet, qui peut déterminer la similarité sémantique entre les mots en calculant la distance entre les nœuds, comme le soulignent les travaux relatifs aux de Liu et Li (2002) et Zhu et Sun (2013). Nous allons utiliser cette similarité pour améliorer la performance du choix des mots dans un *chunk* dans le module du *Semantic Chunk* (les détails vont être présentés lors de la présentation du module).

4.3 Introduction aux réseaux bayésiens

Les réseaux bayésiens (RB) sont apparus à la fin des années 1980 et sont rapidement devenus un des modèles les plus populaires de la communauté IA[38] pour

[38] Le terme « intelligence artificielle », créé par John McCarthy, est souvent abrégé par le sigle « IA » (ou « AI » en anglais, pour Artificial Intelligence). Il est défini par l'un de ses créateurs, Marvin Lee Minsky, comme « la construction de programmes informatiques qui s'adonnent à des tâches qui sont, pour l'instant, accomplies de façon plus satisfaisante par des êtres humains car elles demandent

modéliser l'incertitude. Beaucoup de travaux ont été menés dans ce domaine, comme ceux de Ann et Patrick (1999), Boutilier et al. (1996), d'Avignon et Sauvageau (1994), Éric et Jacques (2007), Jensen (2001), Langley et al. (1992), Sedki et al. (2010), Spirtes et al. (1993) et Stich (2004).

Les réseaux bayésiens sont des modèles graphiques probabilistes représentant des variables aléatoires. Au cours des dernières années, ils sont devenus un outil bien en vogue pour représenter et manipuler des connaissances a priori d'experts dans un domaine d'expertise précis. Ces modèles s'appuient sur le théorème de Bayes, un résultat de base en théorie des probabilités issu des travaux du révérend Thomas Bayes (1702-1761) et présenté à titre posthume en 1763. Nous pouvons utiliser les probabilités pour représenter les phénomènes d'imprécision et d'incertitude.

Le réseau bayésien combine les principes de la théorie des graphes, de la théorie des probabilités, de l'informatique et des statistiques, comme l'ont montré les travaux de Langley et al. (1992) et Fenton et Neil (1999). Ce modèle graphique peut représenter un domaine de connaissance d'une façon intuitive : il est souvent plus facile de formaliser les connaissances sous la forme d'un graphe de causalité que sous la forme d'un système basé sur les règles. En outre, le réseau bayésien peut gérer des ensembles de données incomplètes. De plus, il permet d'identifier la relation causale qui peut nous aider à prendre des décisions.

Un réseau bayésien est composé de deux parties, une qualitative et une quantitative. La partie qualitative est un graphe orienté sans circuit qui peut refléter la structure causale d'un domaine (mais pas toujours). La partie quantitative, quant à elle, représente la distribution de probabilité conjointe des variables. Chaque variable est associée à une table de probabilités conditionnelles (CPT), laquelle représente les probabilités de chaque état de la variable étant donné celui de ses parents. Si une variable n'a pas de variable parent dans le graphe, la table de probabilités

---

des processus mentaux de haut niveau tels que : l'apprentissage perceptuel, l'organisation de la mémoire et le raisonnement critique ».

conditionnelles représente la distribution de probabilité *a priori* de cette variable. Un réseau bayésien est capable de calculer la probabilité a posteriori d'une variable incertaine.

4.3.1 Théorème de Bayes

Voici le théorème de Bayes :

$$P(A|B)P(B) = P(A \cap B) = P(B|A)P(A)$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Le terme $P(A)$ est la *probabilité a priori* de A. Elle est « antérieure » au sens qu'elle précède toute information sur B. $P(A)$ est aussi appelée la *probabilité marginale* de A. Le terme $P(A|B)$ est appelée la *probabilité a posteriori* de A sachant B (ou encore de A *sachant* B). Elle est « postérieure », au sens où elle dépend directement de B. Le terme $P(B|A)$, pour un A connu, est appelée *vraisemblance* de A. De même, le terme $P(B)$ est appelé *probabilité marginale* ou *a priori* de B.

Le réseau bayésien est un modèle qui se fonde sur des hypothèses de Markov pour chercher la satisfaction sur les contraintes d'indépendance. Par exemple, pour déterminer les deux variables X et Y qui sont absolument indépendants, on peut utiliser la probabilité entre elles pour traduire leur indépendance.

Voici la propriété fondamentale des réseaux bayésiens (qui peut être considérée comme leur définition) :

V : ensemble des nœuds

E : ensemble des arcs

G <V, E> : le couple G est un graphe orienté sans circuit (acyclique)

P : une distribution de probabilité jointe sur V

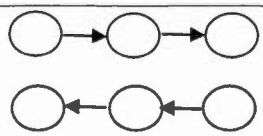
F (V_i) : ensemble des causes (parents) de V_i dans le graphe G

$$P(V_1, V_2, \cdots, V_n) = \prod_{i=1}^{n} P(V_i | F(V_i))$$

Si V_i n'a pas de parent, la probabilité P(Vi) est inconditionnelle. Sinon, elle est conditionnelle.

4.3.2 Modèle graphique du réseau bayésien

Le réseau bayésien est un graphe orienté et sans circuit, donc acyclique. Pour représenter les informations dans un graphe de causalité, nous décrivons au tableau 4.6 l'ensemble des trois événements possibles.

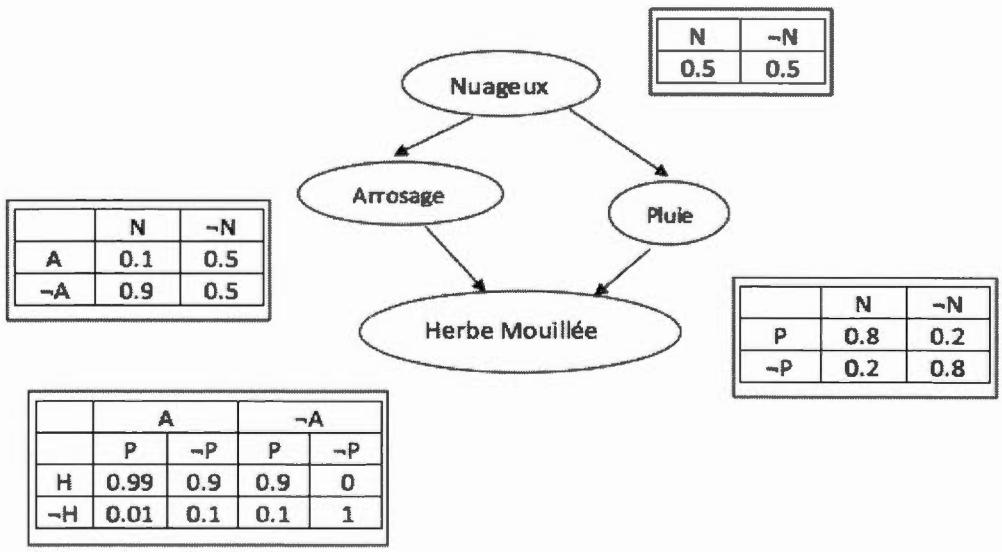
Type	Graphe
Connexion convergente	
Connexion en série	
Connexion divergente	

4.6 : Trois types de connexions

Les réseaux bayésiens sont des réseaux probabilistes basés sur la théorie des graphes. Ils sont composés d'un ensemble de variables et d'un ensemble d'arcs entre ces variables, de sorte que les variables et les arcs forment un graphe dirigé sans circuit. Chaque variable possède un ensemble fini d'états mutuellement exclusifs. À chaque variable A ayant pour parents B_1, ..., B_n est attachée une table de probabilité $P(A|B_1, ..., B_n)$.

L'inférence dans les réseaux bayésiens consiste à calculer P(nœud cible | nœuds observés). Le nœud cible peut-être n'importe où dans le graphe. Les nœuds dont on connaît les valeurs (appelés « observations » ou « évidences ») peuvent aussi se situer n'importe où dans le graphe. L'inférence permet de propager l'information dans le réseau bayésien, c'est-à-dire elle permet de voir comment une nouvelle information vient modifier ce que je "crois" sur une variable.

La figure 4.7 donne un exemple de réseau bayésien.



4.7: Un petit exemple (Ann et Patrick (1999))

Si je ne sais rien, je "crois" qu'il pleut avec une probabilité :

$P(P) = P(P|N) * P(N) + P(P|\neg N) * P(\neg N) = 0.8 * 0.5 + 0.2 * 0.5 = 0.5$

En revanche, si j'observe que le ciel est nuageux, alors je "crois" qu'il pleut avec une probabilité de $P(P|N)= 0.8$.

Voici un autre exemple avec inférence dans le sens inverse d'une flèche : je ne peux pas voir le ciel, mais je peux voir si l'arrosage est en cours ou non. Si j'observe que l'arrosage est en cours, je peux en déduire que le ciel est peu nuageux, car

$P(N|A) = P(A|N) * P(N) / P(A) = P(A|N) * P(N) / (P(A|N) * P(N) + P(A|\neg N) * P(\neg N)) = 0.1 *0.5 / (0.1 * 0.5 + 0.9 * 0.5) = 0.167$. Sans cette observation, je croirais qu'il pleut avec une probabilité de 0.5.

4.4 La méthode de segmentation à base de réseau bayésien

Tel que mentionné précédemment, le module de calcul des probabilités contient plusieurs facteurs. Nous diviserons ces facteurs en deux parties. Une première partie concerne la segmentation du texte hanzi et contient les facteurs qui influencent cette segmentation (par exemple, les symboles, les idiomes, les nouveaux mots, HowNet, etc.). L'autre concerne la segmentation du pinyin, qui contient elle aussi les facteurs influençant cette segmentation (par exemple, la conversion des hanzi aux pinyins, les fautes d'orthographe, etc.). Pour améliorer la performance de la segmentation du texte hanzi, on comparera le résultat de l'analyse des facteurs qui influencent la probabilité de la satisfaction de la segmentation du hanzi avec le résultat de l'analyse des facteurs qui influencent la probabilité de la satisfaction de la segmentation du pinyin. Un exemple similaire a été traité par Fenton et Neil (1999) et Fu et Delcroix (2011) concernant une variété de choix pour un voyage, avec les critères de prix et durée et deux attributs pour les alternatives (moyen de transport et heure de départ).

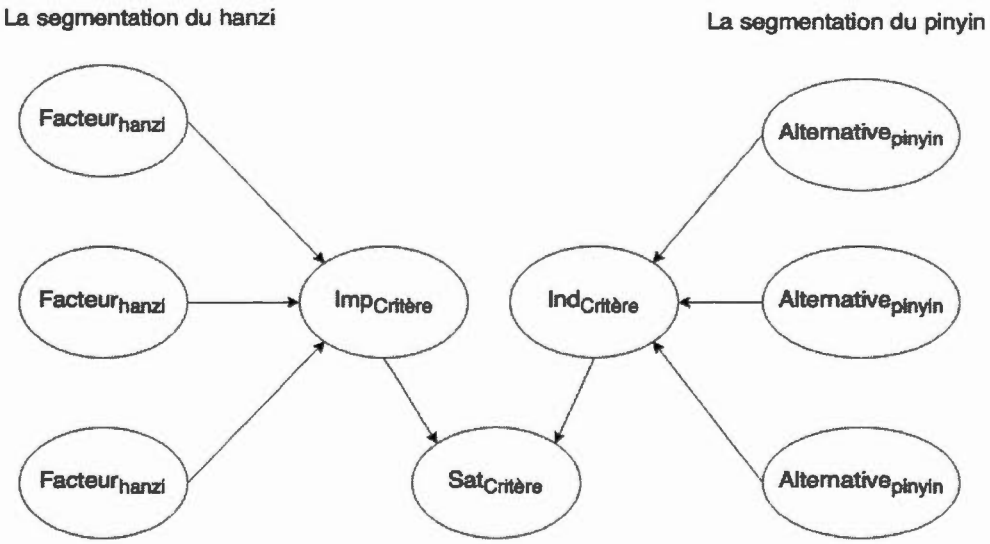
Nous expliquons ci-dessous comment sont définis le graphe et les probabilités du réseau bayésien utilisés pour notre méthode de segmentation, puis nous détaillons les

différentes étapes de notre méthode. Nous définissons d'abord quelques termes du modèle réseau bayésien, tels que présentés dans les travaux de Sedki et al. (2010) :

- Hanzi segmenté: résultat du processus de segmentation du texte chinois. Il est obtenu par le segmenteur de Stanford.
- Pinyin segmenté : résultat du processus de segmentation du pinyin. Il est obtenu par le segmenteur du pinyin[39].
- Alternative : tout facteur pouvant influencer la performance de la segmentation du pinyin. Nous appelons A_{pinyin} représentant l'ensemble des influences sur la segmentation du pinyin, par exemple les fautes d'orthographe, les polyphonies, etc.
- Facteur : tout facteur pouvant influencer la performance de la segmentation du hanzi, par exemple les erreurs syntaxiques, les fautes d'orthographe, les erreurs syntaxiques, la fréquence des mots, etc. Nous appelons F_{hanzi} l'ensemble de ces facteurs.
- Critère : critères déterminant la segmentation du hanzi ou du pinyin, par exemple la structure de la phrase, le mot clé et le *Semantic Chunk.*
- Indice d'un critère : fonction numérique sur l'ensemble des alternatives qui représentent la qualité d'une alternative pour ce critère. Nous l'appelons $Ind_{Critère}$.
- Importance d'un critère : fonction numérique qui représente le niveau d'importance d'un critère pour un texte segmenté donné (en fonction de ses influences). Nous l'appelons $Imp_{Critère}$.
- Satisfaction d'un facteur : niveau de satisfaction de la segmentation du texte sur une alternative pour un critère. Nous l'appelons $Sat_{Critère}$.

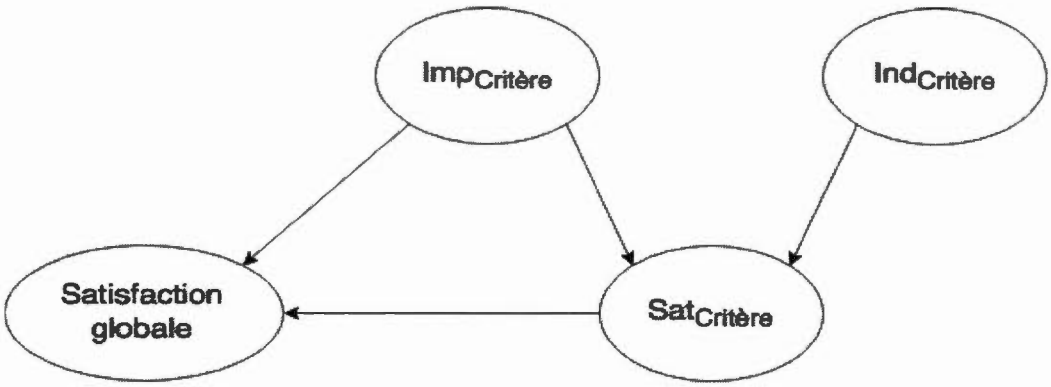
[39] https://github.com/hotoo/pinyin

4.5 La structure générale du graphe du réseau bayésien



4.8 : Graphe du réseau bayésien associé à un facteur

Les réseaux bayésiens associés à chaque facteur sont regroupés pour constituer le réseau bayésien du problème complet. Puis on ajoute un nœud *satisfaction globale* qui dépend des nœuds *satisfactions* et *importances* de chaque facteur.



4.9 : Graphe du réseau bayésien associé le nœud de la satisfaction globale

C'est le modèle que nous proposons. Il peut calculer une valeur de satisfaction entre la segmentation du hanzi et du pinyin, c'est-à-dire identifier la meilleure correspondance entre le hanzi et le pinyin en évitant le problème des mauvaises entrées des hanzi, car il peut identifier la polyphonie.

4.5.1 Valeurs des trois types de noeuds

Les trois types de nœuds $Imp_{Critère}$, $Ind_{Critère}$, $Sat_{Critère}$ ont une valeur dans l'intervalle [0,1].

Valeur	Interprétation		
	$Imp_{Critère}$	$Ind_{Critère}$	$Sat_{Critère}$
0 – 0.2	indifférent	très mauvais	très mauvais
0.2 – 0.4	peu important	mauvais	mauvais
0.4 – 0.6	moyen	moyen	moyen
0.6 – 0.8	important	bon	Bon
0.8 - 1	très important	très bon	très bon

4.10 : La valeur de nœuds $Imp_{Critère}$, $Ind_{Critère}$, $Sat_{Critère}$

4.5.2 Les facteurs d'influence de la segmentation du hanzi

Nous avons choisi 3 importances, chacune s'associant avec différents facteurs de la segmentation du texte. Nous expliquons ci-dessous la fonction des différents facteurs qui influencent la performance de la segmentation du hanzi.

Les critères :

- ***C_{n_d} (Facteur nouveau-dictionnaire)***. Nous avons déjà présenté la façon dont HowNet va interagir avec le grand dictionnaire pour en donner une forme regroupée. Nous allons utiliser cette forme regroupée comme un facteur qui influence la segmentation.
- ***C_i (Facteur dictionnaire des idiomes).*** Les idiomes en chinois sont souvent formés de quatre hanzi. Le segmenteur de Stanford divise toujours les idiomes en deux parties de deux hanzi. Ce facteur peut augmenter la probabilité entre ces deux parties pouvant donc déterminer à la fin du processus de segmentation que ces deux parties forment un seul mot (idiome).

- C_m ***(Facteur dictionnaire mis à jour).*** À la fin de chaque processus de segmentation, le système va systématiquement mettre à jour les nouveaux mots ou les nouvelles expressions dans un dictionnaire mis à jour. Ces mots ou expressions pourront donc être réutilisés lors de la prochaine segmentation.
- C_g ***(Facteur grammaire).*** Des erreurs de POS tagging ou d'autres fautes dans l'étiquetage grammatical vont influencer les probabilités de la segmentation. Ce facteur va corriger ces erreurs.
- C_{s_c} ***(Facteur sémantique chunk).*** Un « sémantique chunk » tel qu'il ressort des travaux de Zhang et Simon (1985), Wen et al. (2014), Yang et Zong (2014) (Zhang, Sun, & Zhou, 2012), Zhou et al. (2002), Sha et al. (2012), est un groupe de commentaires consécutifs (sur une plateforme de réseaux sociaux) qui portent autour d'un même thème. Ce facteur contient deux parties; l'une concerne la fréquence des mots contenus dans ce *chunk*, et l'autre partie désigne la similarité des mots de ce *chunk*. Nous utilisons dans ce cas, HowNet pour calculer la distance entre les mots du *Semantic Chunk* (qui sera identique à la distance de ces mots dans le réseau HowNet). Les mots les plus fréquents ou les plus semblables dans ce *chunk* vont avoir plus d'influence sur les probabilités de la segmentation.
- C_f ***(Facteur de la fréquence des mots du grand corpus).*** Il s'agit ici de la fréquence des mots dans le corpus entier, qui est différente de la fréquence des mots dans le *Semantic Chunk*.
- C_s ***(Facteur symbole).*** Les symboles sont beaucoup utilisés dans les médias sociaux, mais peu ou pas du tout dans les textes plus traditionnels, comme les journaux, les livres, etc. Dans le corpus de commentaires provenant des médias sociaux, les emojis[40] sont beaucoup utilisés; certains représentent simplement des émotions, mais d'autres peuvent représenter des mots. Les

40 https://fr.wikipedia.org/wiki/Emoji.

emojis sont enregistrés comme code ASCII dans le fichier HTML. Au cours du processus de segmentation, nous pouvons donc convertir ces codes en hanzi ou en pinyin; cette conversion va influencer la probabilité d'une segmentation uniquement du corpus des médias sociaux, et non pour le corpus en général.

Ces sept facteurs ci-dessus énumérés vont donc fortement influencer les nœuds $Importance_{Facteur}$.

4.5.3 Les nœuds de l'Importance

Comme nous l'avons décrit auparavant, ces facteurs vont influencer la segmentation du hanzi de différentes façons. Nous considérons ici les trois différentes Importances des Critères. Chaque critère combine différents facteurs.

- ***Importance_StructurePhrase.*** Ce nœud représente la probabilité de l'importance du critère « structure de la phrase » sur ses facteurs associés. Les facteurs du nouveau-dictionnaire (C_{n_d}), de la grammaire (C_g), du dictionnaire mis à jour (C_m), du dictionnaire idiomes (C_i) et du symbole (C_s) vont tous influencer cette Importance. Si la probabilité de cette Importance est plus proche de 1, cela signifie que la structure grammaticale de cette phrase est presque juste. Si la probabilité est plus proche de 0, cela signifie alors que la structure grammaticale de cette phrase est peut-être défaillante. Cette Importance variera selon le corpus : dans les commentaires tirés des médias sociaux, il y a beaucoup de phrases qui n'ont pas de structure complète. Aussi, pour ce corpus, les deux facteurs C_{n_d} et C_g auront d'influence sur la valeur de la segmentation tout en réduisant l'impact de cette Importance. À l'inverse, les deux facteurs C_{n_d} et C_g seront beaucoup plus déterminants pour la construction de la phrase si le système analyse un corpus formel, par exemple

des journaux ou des documents spécialisés dont la grammaire est plus rigoureuse.

- ***Importance_MotCle.*** Ce nœud représente la probabilité de l'importance du critère « mots clés » sur ses facteurs associés. Cette probabilité ne dépend pas que de la fréquence des mots (C_f), mais aussi des facteurs nouveau-dictionnaire (C_{n_d}), symbole (C_s) et dictionnaire mis à jour (C_m). Le facteur de la fréquence des mots et celui du nouveau dictionnaire ont la même importance dans les deux corpus (médias sociaux et formels). La probabilité de ce nœud ne change pas selon les corpus. Puisque, contrairement aux corpus formels, le corpus tiré des médias sociaux contient beaucoup d'emojis, nous pouvons donc augmenter l'impact de ce facteur pour cette Importance au sein du corpus tiré des médias sociaux, et en réduire l'impact dans le corpus formel. Si la probabilité de ce nœud est plus proche de 1, cela signifie que ces mots ou symboles sont plus fréquents dans le corpus, et donc qu'ils peuvent être considérés comme étant le mot clé du corpus. Toutefois, dans un tel cas, nous considérons aussi les mots fonctionnels et les mots d'arrêt (*stop words*). S'ils apparaissent souvent à une grande fréquence dans le corpus, ces mots ne peuvent pas être considérés comme étant le mot clé du corpus. En revanche, si la probabilité est plus proche de 0, cela signifie que ces mots apparaissent très rarement dans le corpus et donc qu'ils ont une petite probabilité d'être un mot clé.
- ***Importance_SemanticChunk.*** Ce nœud représente la probabilité de l'importance du critère *Semantic Chunk* sur ses facteurs associés. Cette probabilité dépend des facteurs de fréquence des mots du grand corpus (C_f) et du *Semantic Chunk* (C_{s_c}). Si la probabilité est plus proche de 1, cela signifie que ces mots sont plus importants pour le grand corpus et aussi pour le *chunk*. Ici, nous considérons aussi les mots fonctionnels et les mots d'arrêt comme nous le faisions pour le nœud Importance_MotCle. Si, en revanche, la

probabilité est plus proche de 0, cela signifie que ces mots ne sont importants ni pour le grand corpus, ni pour le *chunk*. Si la probabilité est autour de 0,5, cela signifie que ces mots sont probablement plus importants pour le grand corpus que pour ce *chunk,* ou inversement, qu'ils sont plus importants pour ce *chunk* que pour le grand corpus.

4.6 Les facteurs d'influence de la segmentation du pinyin

La partie droite du réseau bayésien concerne les facteurs de l'influence pour la segmentation du pinyin. Nous considérons ici ces différents facteurs :

- A_s ***(Facteur symboles)***. Ce facteur fonctionne comme nous l'avons décrit dans la partie sur la segmentation du hanzi.
- ***Ag (Facteur grammaire).*** Après le processus de la segmentation du pinyin, la phrase peut présenter des problèmes grammaticaux.
- A_n ***(Facteur nouveau-structure)***. Dans la partie relative à la problématique de la segmentation, nous avons vu qu'il existe de nouvelles structures : hanzi (pinyin) hanzi (pinyin). Ce facteur peut augmenter la probabilité entre le groupe des deux hanzi et pareillement pour les groupes de deux pinyins, ou inversement, en diminuer la probabilité entre le hanzi et le pinyin, et le pinyin et le hanzi.
- A_m ***(Facteur dictionnaire mis à jour)***. Ce facteur a une fonction identique à celle présentée pour la segmentation du hanzi.
- A_f ***(Facteur fréquence des pinyins)***. Ce facteur va calculer la fréquence de chaque pinyin du grand corpus.
- A_{s_c} ***(Facteur semantic chunk)***. Ce facteur va calculer la fréquence de chaque pinyin du *chunk*. Il fait la même chose que le côté de la segmentation du hanzi.

- ***A_p (Facteur polyphonie)***. Dans le système linguistique chinois, il existe beaucoup de polyphonie. Ce facteur va influencer le choix des hanzi.
- ***Ah (Facteur choix-hâtif).*** Pendant le processus de choix du hanzi correspondant au pinyin, un utilisateur peut faire hâtivement un mauvais choix. Nous allons donc augmenter ou baisser la probabilité quand le hanzi apparaît.
- ***A_d (Facteur dictionnaire-pinyin)***. Contrairement au dictionnaire hanzi, le dictionnaire pinyin ne se combine pas avec le système HowNet. Il s'agit donc simplement d'un vaste dictionnaire qui associe les hanzi au(x) pinyin(s) correspondants.

4.7 Les nœuds de l'Indice

Comme les nœuds de l'*Importance*, ici nous considérons aussi les trois nœuds de l'*Indice* correspondants.

- ***Indice_StructurePhrase.*** Ce nœud représente l'ensemble des facteurs portant sur la qualité de la structure de la phrase. Cet indice dépend des facteurs du dictionnaire pinyin (A_d), du symbole (A_s), des nouvelles structures (A_n), du dictionnaire mis à jour (A_m) et de la grammaire (A_g). Tous ces facteurs vont faire augmenter ou baisser les probabilités du pinyin de la phrase.
- ***Indice_MotCle.*** Nous utilisons ici le MotCle pour correspondre au nœud Importance_MotCle, mais « mot » ici représente le pinyin, et non pas le sinogramme. La probabilité de ce nœud dépend des facteurs de la fréquence des pinyin du grand corpus (A_f), du dictionnaire pinyin (A_d), du polyphone (A_p) et du choix hâtif (A_h). Si la probabilité de cet indice est plus proche de 1, cela signifie qu'il est très probable que le hanzi soit bien identifié et qu'il apparaisse souvent dans le grand corpus. Si la probabilité de ce nœud est plus

proche de 0, cela signifie que ce pinyin est un polyphone, et donc que le hanzi sera difficilement identifié et il n'apparaîtra pas dans le grand corpus.

- ***Indice_SemanticChunk.*** Ce dernier nœud dépend des facteurs du *Semantic Chunk* (A_{s_c}) et de la fréquence du pinyin (A_f). La probabilité associée à ce nœud est calculée de la même façon que la probabilité du nœud Importance_SemanticChunk. Si elle est plus proche de 1, cela signifie que ce pinyin apparaît plus souvent dans le grand corpus et aussi dans le *chunk*. Au contraire, si la probabilité est proche de 0, cela signifie que ce pinyin apparaît rarement dans le grand corpus et le *chunk*.

4.8 Les nœuds de Satisfaction

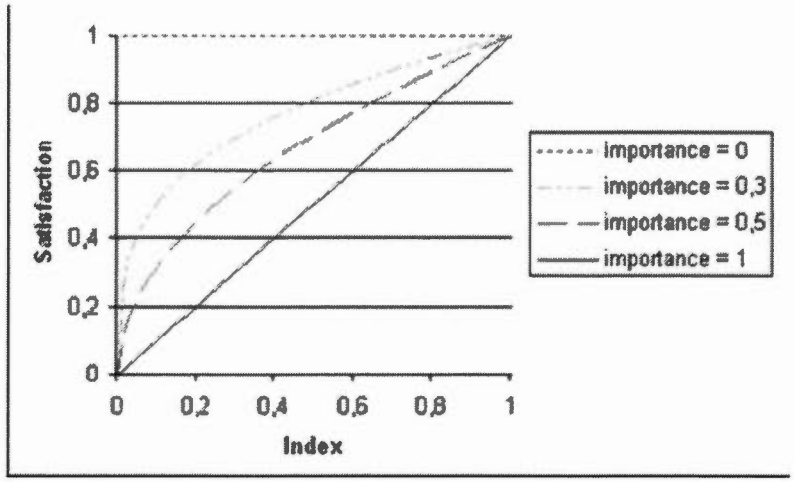
Les trois premiers nœuds de *Satisfaction* décrivent la probabilité de la satisfaction entre les nœuds *Importance* et les nœuds *Indice*.

- ***Satisfaction_StructurePhrase.*** Ce nœud indique une similarité entre la structure de la phrase écrite en hanzi et celle écrite en pinyin. Si la valeur du nœud est plus proche de 1, cela signifie que les deux structures concordent. En revanche, si la valeur est plus proche de 0, cela signifie que la structure de la phrase écrite en hanzi ne correspond pas à celle de la phrase écrite en pinyin. Dans ce cas, il existe des transcriptions de pinyin segmentées qui ne correspondent pas aux mots segmentés en hanzi, et par conséquent, il existe des différences possibles de segmentation entre les textes en hanzi et en pinyin.
- ***Satisfaction_MotCle.*** Ce nœud représente la similarité entre les mots clés dans les deux corpus segmentés.
- ***Satisfaction_SemanticChunk.*** Ce nœud représente la similarité entre le *Semantic Chunk* en pinyin et celui hanzi. Il comprend les mots clés et la fréquence des mots des *chunks*.

- ***Satisfaction_segmentation.*** Ce dernier nœud décrit la qualité de la segmentation des hanzi et des pinyins. Si sa probabilité est plus proche de 1, cela signifie que les deux segmentations sont similaires, que les mots segmentés ou les transcriptions du pinyin segmenté sont semblables, avec un résultat de segmentation probant, soit juste et précis. En revanche, si la probabilité de ce nœud est plus proche de 0, cela signifie que les phrases segmentées des deux côtés sont distinctes : des mots hanzi ne correspondent pas aux pinyins, et inversement. Dans ce cas, le texte contient beaucoup de mots OOV qui ne sont pas mis à jour dans le dictionnaire.

Les variables de SatCritère représentent le degré de satisfaction de la segmentation du hanzi en correspondance avec la segmentation du pinyin :

$$Sat(c) = Ind(c)^{Imp(c)}$$



4.11 : Satisfaction d'un critère en fonction de l'indice de qualité d'une alternative

La satisfaction est constante, égale à 1 lorsque le critère est indifférent pour la segmentation du hanzi (ce critère peut être ignoré). Dans le cas contraire, la

satisfaction augmente avec la qualité de l'alternative, et pour une alternative donnée, la satisfaction est plus élevée lorsque le critère est peu important.

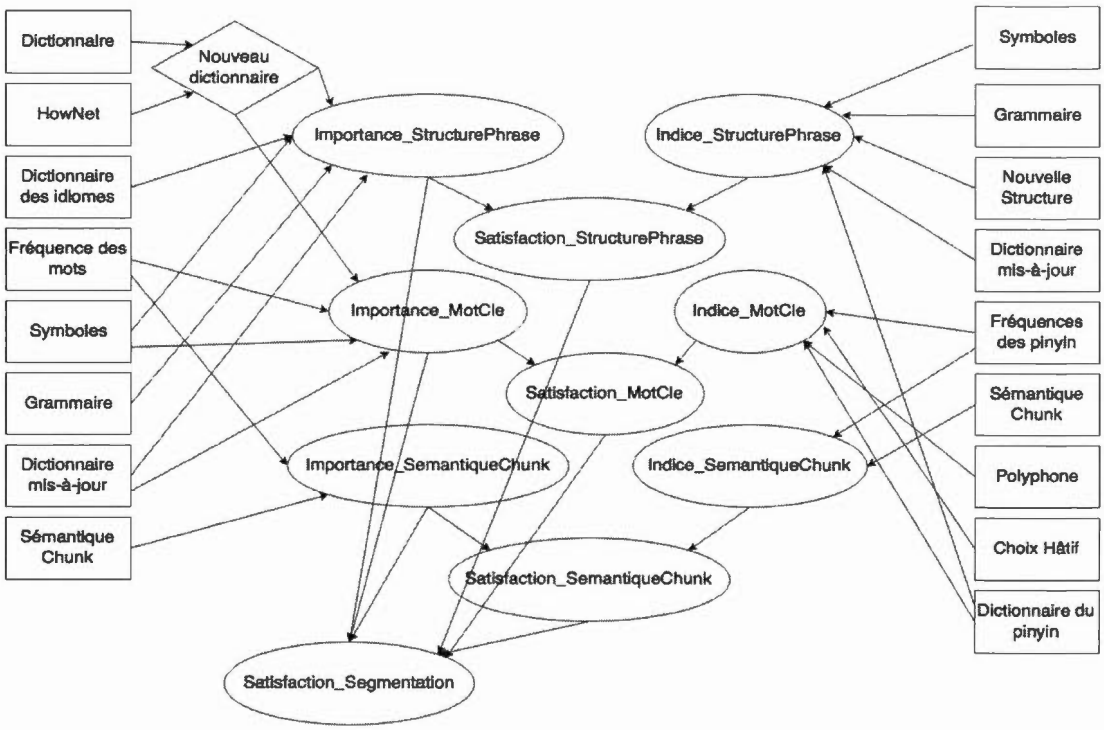
Cette fonction permet de définir automatiquement la table de probabilités conditionnelles du nœud de satisfaction.

La table de probabilités du nœud "satisfaction_Segmentation" est définie par la fonction :

$$Satisfaction\_Segmentation = \frac{\sum_i \mathrm{ImpC_i} * \mathrm{SatC_i}}{\sum_i \mathrm{ImpC_i}}$$

Une fois le réseau bayésien (RB) entièrement défini (graphe et probabilités), il peut être utilisé pour calculer les probabilités *a posteriori* pour les attributs des alternatives.

Le graphe complet du système à la figure 4.12 démontre les différents facteurs, les nœuds de l'Importance et de l'Indice, ainsi que les nœuds de Satisfaction dépendant des trois types de nœuds de l'Importance et de l'Indice.



4.12 : Le graphe du système RB complet

4.9 Le système complet proposé

La plupart des outils de segmentation fonctionnent assez bien pour les textes généraux, comme les textes journalistiques, ceux provenant de magazines et les textes officiels, etc., mais ils ne fonctionnent pas très bien pour les textes spécifiques à un domaine d'expertise, tel que les textes médicaux ou de chimie, ainsi que pour les textes des médias sociaux, comme Weibo ou Facebook (i.e., les commentaires en chinois dans Facebook), etc. Nous présentons ici le taux de réussite de trois différents outils de segmentation : 1) le segmenteur de Stanford en version 2015, qui intègre un dictionnaire contenant 423 200 mots; 2) le segmenteur *Natural Language Processing & Information Retrieval Sharing Platform, Institute of Computing Technology,*

Chinese Lexical Analysis System (NLPIR-ICTCLAS) en version 2014, qui est développé par des chercheurs du Institute of Computing Technology (Chine) ; 3) le segmenteur du *Language Technology Platform Cloud* (LTP-Cloud) en version 2014, du Harbin Institute of Technology (Chine)[41].

Nous avons utilisé deux types de corpus pour faire le test : un corpus général tiré d'un journal, et un autre tiré de Sina Weibo (un microblogue très populaire en Chine, qui est un mélange de Facebook et de Twitter), lequel contient de nouveaux mots, de nouvelles expressions, du pinyin, des symboles et de nouveaux idiomes. La statistique du corpus général figure aux tableaux 4.13 et 4.14.

	Hanzi	Mot	Expression	Idiome	Pinyin	OOV	Symbole
Nombre	2216	1256	27	14	0	36	0
% par rapport le mot	-	-	2.15%	1.11%	0%	2,87%	0%

4.13 : Statistique du nombre de différents types de mots

Le corpus analysé dans ce tableau contenait au total 2 216 hanzi et 1 256 mots. Nous avons analysé les résultats selon six aspects : les mots, les expressions, les idiomes, le pinyin, les nouveaux mots (OOV) et les symboles. Voici d'abord neuf

41 Ces deux derniers segmenteurs sont des systèmes fermés. Nous n'avons pas d'information sur le nombre de mots que contient leur dictionnaire.

définitions des attributs qui reflètent la performance des trois outils de segmentation. Les définitions de la précision, du rappel et du F-mesure sont décrites par[42].

- Nombre : il s'agit du nombre total de mots donnés par la segmentation. Ce nombre n'est pas le nombre de mots du texte, car il contient aussi les mots mal segmentés. Par exemple, dans le texte segmenté 撤 | 地 | 设 | 市 (che4di4she4shi4) quatre mots sont comptés, mais le nombre correct est en fait un seul mot43.
- Précision[44]: « *An information retrieval performance measure that quantifies the fraction of retrieved documents which are known to be relevant.* » C'est la proportion du nombre de mots correctement segmentés par rapport au nombre total de mots segmentés.

$$Précision = \frac{Le\ nombre\ de\ mots\ correctement\ segmentés}{Le\ nombre\ total\ de\ mots\ segmentés}$$

- Rappel: « *An information retrieval performance measure that quantifies the fraction of known relevant documents which were effectively retrieved.* » C'est la proportion du nombre de mots correctement segmentés par rapport au nombre total de mots connus.

$$Rappel = \frac{Le\ nombre\ de\ mots\ correctement\ segmentés}{Le\ nombre\ total\ de\ mots\ connus}$$

---

[42] Baeza-Yates, Ricardo; Ribeiro-Neto, Berthier (1999). Modern Information Retrieval. New York, NY: ACM Press, Addison-Wesley, Seiten 75 ff.

[43] Ici, nous utilisons le training date : http://www.icl.pku.edu.cn/icl_groups/corpus/coprus-annotation.htm pour exploiter le corpus des médias sociaux et http://sighan.cs.uchicago.edu pour le corpus plus général, comme celui du milieu journalistique.

[44] https://en.wikipedia.org/wiki/Precision_and_recall

		Mot	Expression	Idiome	OOV
Stanford Segmenteur	Nombre	1463	7	4	2
	Précision	0,714	-	-	-
	Rappel	0,831	0,185	0,214	0
	F-mesure	0,768	-	-	-
	Riv	0,856	-	-	-
NLPIR (ICTCLAS)	Nombre	1245	10	5	4
	Précision	0,793	-	-	-
	Rappel	0,786	0,296	0,357	0,028
	F-mesure	0,789	-	-	-
	Riv	0,809	-	-	-
LTP-Cloud	Nombre	1197	13	4	5
	Précision	0,831	-	-	-
	Rappel	0,792	0,407	0,286	0,028
	F-mesure	0,811	-	-	-
	Riv	0,816	-	-	-

4.14 : Statistique de la segmentation effectuée par les trois outils pour le texte général

- F-Mesure[45]: « *A measure that combines precision and recall is the harmonic mean_of precision and recall, the traditional F-measure or balanced F-score.*» Elle établit un équilibre entre la précision et le rappel qui décrit la performance du système.

$$F - score = \frac{2 * Précision * Rappel}{Précision + Rappel}$$

- Le taux de mots « hors vocabulaire » (taux dOut-Of-Vocabulary (OOV)): C'est la proportion du nombre de mots inconnus par rapport au nombre total de mots du corpus général.

$$OOV = \frac{Le\ nombre\ de\ mots\ inconnus}{Le\ nombre\ total\ de\ mots\ du\ corpus}$$

- Rappel sur Expression : C'est la proportion du nombre d'expressions correctement segmentées par rapport au nombre total d'expressions.

$$Rexpression = \frac{Le\ nombre\ d'expression\ correctement\ segmentés}{Le\ nombre\ total\ d'expressions}$$

- Rappel sur Idiome : C'est la proportion du nombre d'idiomes correctement segmentés par rapport au nombre total d'idiomes.

$$Ridiome = \frac{Le\ nombre\ d'idiomes\ correctement\ segmentés}{Le\ nombre\ total\ d'idiomes}$$

- Rappel sur OOV : C'est la proportion du nombre de mots inconnus correctement segmentés par rapport au nombre total de mots inconnus.

$$Roov = \frac{Le\ nombre\ de\ mots\ inconnus\ correctement\ segmentés}{Le\ nombre\ total\ de\ mots\ inconnus}$$

[45] https://en.wikipedia.org/wiki/F1_score

- Rappel sur les mots « dans le vocabulaire » (*Recall on in-vocabulary [Riv]*): C'est la proportion du nombre de mots connus correctement segmentés par rapport au nombre total de mots connus.

$$Riv = \frac{Le\ nombre\ des\ mots\ connus\ correctement\ segmentés}{Le\ nombre\ total\ de\ mots\ connus}$$

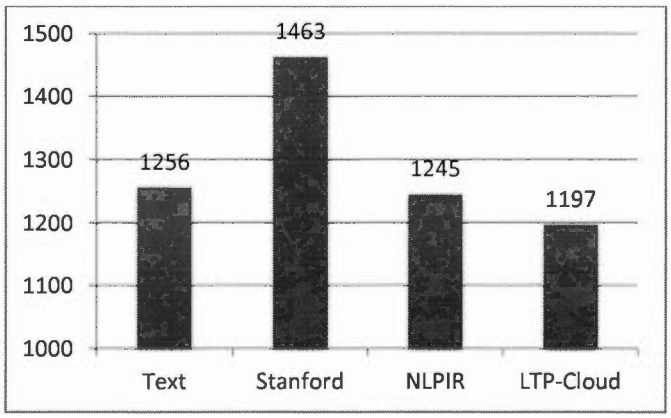
Pour le premier traitement, nous avons choisi un petit texte journalistique, soit la page 30 du *People's Daily* du 20 novembre 2014, qui contient 2 216 hanzi. Le segmenteur de Stanford considère que le corpus contient 1 463 mots; NLPIR qu'il y en a 1 245 mots; et 1 197 mots pour LTP-Cloud. Comme nous le montrons au tableau 2, les trois outils ont fait de mauvaises segmentations. Les erreurs sont résumées ci-après. Nous précisons que tous les exemples cités proviennent du corpus étudié.

1. Les noms propres : les noms propres posent un grand problème pour le processus de segmentation. Il contient le nom des personnes, des villes, des organisations, des concours, etc. Par exemple : 和县官亭镇 (he2xian4guan1ting2zhen4, le nom d'un petit village) que le segmenteur de Stanford segmente ainsi 和|县官|亭|镇.
2. Les nouvelles expressions : une expression est un ensemble de caractères servant à désigner une chose, mais pour laquelle il n'existe pas de forme unique. L'expression n'est donc pas considérée comme un idiome, mais les gens l'utilisent néanmoins souvent, les caractères sont donc souvent combinés ensemble. Par exemple, les trois outils ont segmenté le mot 撤地设市 (che4di4she4shi4) ainsi 撤|地|设|市. Cette expression est utilisée par le gouvernement chinois pour décrire l'attribution du statut administratif spécial de district pour une grande ville, et le résultat ne devrait pas être segmenté.

3. Les nouveaux idiomes : comme mentionné précédemment, un idiome existant peut être changé en modifiant un ou deux caractères, créant ainsi un nouvel idiome.
4. « Une région de ... » : Cette formule est toujours utilisée pour décrire une région ou une combinaison de régions. Par exemple : 工业园 (gong1ye4yuan2, le parc industriel).
5. Un idiome existant est divisé en deux mots au milieu; par exemple : AA|AA (rappelons que les idiomes chinois sont habituellement composés de quatre caractères). Parce que le dictionnaire est incomplet, le segmenteur de Stanford divise toujours l'idiome en deux mots et en son milieu. Cette segmentation brise la signification complète de l'idiome.
6. La ponctuation « " » et « " » : les guillemets ont deux fonctions dans le texte chinois. La première sert à exprimer la parole d'une personne. Le contenu des guillemets peut alors être segmenté séparément. La deuxième fonction est pour décrire une chose spécifique, une expression ou un énoncé particulier. Le contenu des guillemets ne peut alors être segmenté. Par exemple, on retrouve des guillemets avant et après "两化"(liang3hua4). Ce mot est utilisé pour décrire l'industrialisation informatisée (comme une chaîne de montage contrôlée par ordinateur). En chinois, industrialisation est 工业化 (gong1ye4hua4) et ce qui est informatisé est 信息化 (xin4xi1hua4). Dans l'expression entre guillemets, le premier caractère signifie « deux » 两 (liang3) et le deuxième caractère est le même caractère que le troisième caractère d'industrialisation et informatisé soit 化 (hua4), une particule signifiant « isation » ou « rendre » (comme dans « rendre industriel » et « rendre informatique »). Nous pouvons considérer ce contenu, qui signifie donc littéralement quelque chose comme « les deux 'isations' », comme une forme d'abréviation ; les guillemets servent à marquer ce statut. Le contenu interne des guillemets ne peut donc pas, dans ce cas être segmenté.

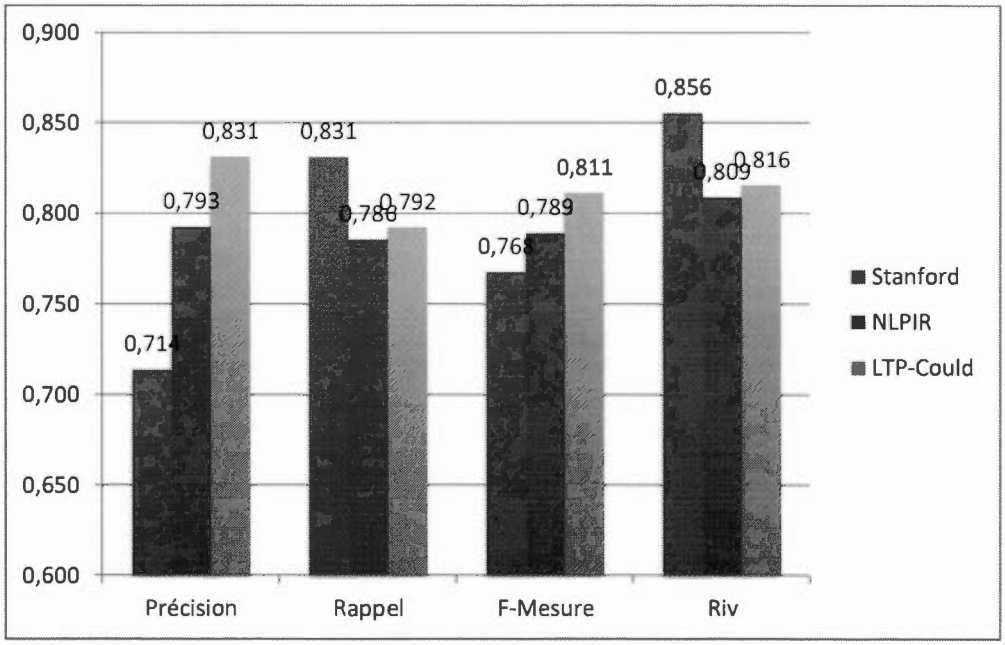
7. La ponctuation « 、 » : Il s'agit d'une ponctuation spéciale du texte chinois. Elle est utilisée pour séparer les locutions juxtaposées. Par exemple : 黄色、绿色、红色和蓝色 (huang2se4, lu4se4, hong2se4he2lan2se4). Cette ponctuation fonctionne comme la virgule dans la phrase « jaune soleil, vert pomme, rouge pompier et bleu ciel ». Nous ne pouvons pas segmenter les parties entre les deux ponctuations sans perdre le sens des mots (dans l'exemple en français, le texte ne parle pas de soleil, de pommes, etc., mais plutôt de couleurs).

La plupart des types d'erreurs de segmentation sont présentés ci-dessus. Les figures 4.15, 4.16 et 4.17 ci-dessous dévoilent le résultat de la performance des trois outils.



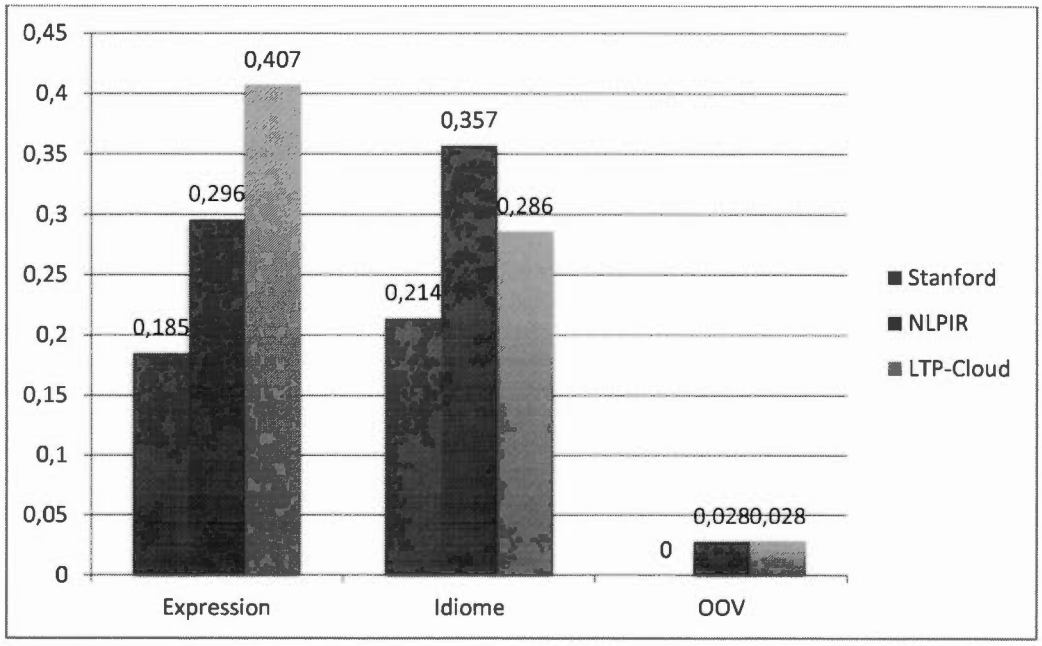
4.15 : Nombre de mots sur le petit corpus du journal

Selon la figure 4.15, le nombre de mots segmentés par les trois types d'outils est près du nombre de mots identifiés avec exactitude. Les trois outils présentent différents résultats de segmentation, lesquels dépendent de leurs stratégies de segmentation respectives ou du nombre de mots de leur dictionnaire intégré.



4.16 : Statistique de la performance sur le petit corpus du journal

Dans la figure 4.16, on voit que la valeur de F-score de LTP-Cloud est plus élevée (0,811) que celle des deux autres outils, et donc la performance de LTP-Cloud est meilleure que les deux autres. Le segmenteur de Stanford a une valeur de Riv (0,856) plus élevée que les deux autres outils, c'est-à-dire que le pourcentage de mots connus est plus élevé que les autres, et nous pouvons donc présumer que le dictionnaire intégré du Stanford contient plus de mots que les deux autres. La performance moyenne de LTP-Cloud est meilleure que celle des deux autres.



4.17 : Statistique des rappels de différents tris de mots sur le petit corpus du journal

Selon la figure 4.17, le segmenteur de Stanford a une mauvaise performance pour les expressions, les idiomes et les OOV. NLPIR, en revanche, fait un bon travail pour segmenter les OOV et les idiomes, quant au LTP-Cloud, le résultat est probant pour segmenter les expressions.

Voici une deuxième analyse, cette fois relative au corpus tiré du média social Sina Weibo. Il contient 2 357 hanzi et 1 578 mots. Les trois outils utilisés sont les mêmes que lors de la première analyse.

	Hanzi	Mot	Expression	Idiome	Pinyin	OOV	Symbole
Nombre	2357	1578	10	9	14	57	38
% par rapport le mot	-	-	2.15%	1.11%	0%	2,87%	0%

4.18 : Statistique du nombre de différents tris de mots du microblogue, Sina Weibo

		Corpus microblogue					
		Mot	Expression	Idiome	Pinyin	OOV	Symbole
Stanford Segmenteur	Nombre	1876	5	5	3	14	9
	Précision	0,563	-	-	-	-	-
	Rappel	0,670	0,200	0,222	0	0,104	0

		NLPIR (ICTCLAS)						LTP-Cloud	
	F-mesure	Riv	Nombre	Précision	Rappel	F-mesure	Riv	Nombre	Précision
	0,611	0,720	1623	0,683	0,703	0,693	0,723	1425	0,781
	-	-	8	-	0.400	-	-	7	-
	-	-	4	-	0,111	-	-	6	-
	-	-	3	-	0	-	-	3	-
	-	-	17	-	0,158	-	-	21	-
	-	-	11	-	0	-	-	11	-

	Rappel	0,804	0,300	0,222	0	0,211	0
	F-mesure	0,793	-	-	-	-	-
	Riv	0,826	-	-	-	-	-

4.19 : Précision de la segmentation par les trois outils pour le texte de microblogue, Sina Weibo

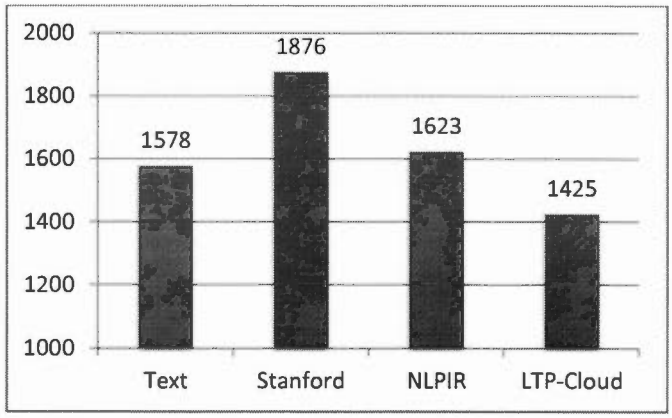
Pendant le processus de segmentation du texte des médias sociaux, nous avons relevé divers problèmes reliés aux segmenteurs, dont voici le résumé :

1. La ponctuation « “ » et « ” » : ce problème est identique à celui que nous avions rencontré dans le corpus général.
2. Le caractère typographique « # » … « # » : dans les textes tirés des microblogues, le contenu entre-deux « # » indique un sujet ou un thème. Il existe un espace avant et après le « # ». Lorsque le texte se présente sous cette forme, on peut considérer que le contenu entre ces caractères typographiques est une nouvelle expression ou un nouveau mot. Par exemple : #小时代# (xiao3shi2dai4). LTP-Cloud propose une segmentation en 小|时代 (petit | époque). Cependant, il s'agit là d'un nom de film, et nous pouvons donc considérer que c'est une expression unique qui contient trois caractères.
3. Le caractère typographique « @ » : L'usage de ce caractère typographique dans les microblogues est « un espace + @ + un/plusieurs caractère/s + un espace ». Le texte mis sous cette forme représente le fait qu'une personne *fait*

référence à ou *appelle* une autre personne. Le contenu entre « @ » et l'espace à la fin désigne le nom de la personne dans le microblogue. On devra faire attention de distinguer l'usage de cette ponctuation avec la forme de l'adresse courriel, comme ***@***.com, ***@***.ca, ***@***.edu, etc.

4. La ponctuation relative aux sites web : la forme « http:// » dans les textes indique un lien vers un site web ; il n'est donc pas nécessaire de faire la segmentation entre les lettres lorsqu'on trouve ce type de ponctuation.
5. Les mauvais caractères et les bons caractères ayant un pinyin identique: au cours du processus d'entrée des caractères chinois par le clavier, les utilisateurs peuvent faire des erreurs de choix, comme nous l'avons mentionné précédemment. Si en entrant le texte, l'outil fait une segmentation entre deux caractères, cela veut dire que ces deux caractères ont une petite probabilité de former un mot (nous ne considérons ici que le pinyin de deux caractères et pour lequel les deux transcriptions de pinyin jointes bout à bout forment un mot). Prenons par exemple les expressions 麦咖啡 (mai4ka1fei1, blé café) et 卖咖啡 (mai4ka1fei1, vendre café). Le caractère 卖 (mai4, vendre) est un verbe, mais le caractère différent 麦 (mai4, blé), qui est homophonique avec le précédent, est un nom ou, ici, un adjectif pour décrire un type de café (le café de blé). C'est le fait que ces deux caractères — dont les fonctions grammaticales sont différentes — ont le même pinyin qui a causé le problème grammatical.
6. Les abréviations : l'entrée de texte sur les microblogues chinois est limitée à 140 hanzi; les utilisateurs emploient donc beaucoup d'abréviations. Par exemple, le mot 个算 (ge4suan4) n'a pas de signification, donc nous avons besoin de le segmenter en 个|算. Le premier hanzi est l'abréviation du mot 这个 (zhe4ge4, celui, celle), le second est 计算 (ji4suan4, calculer), et donc l'abréviation précédente signifie « cette calculatrice ».

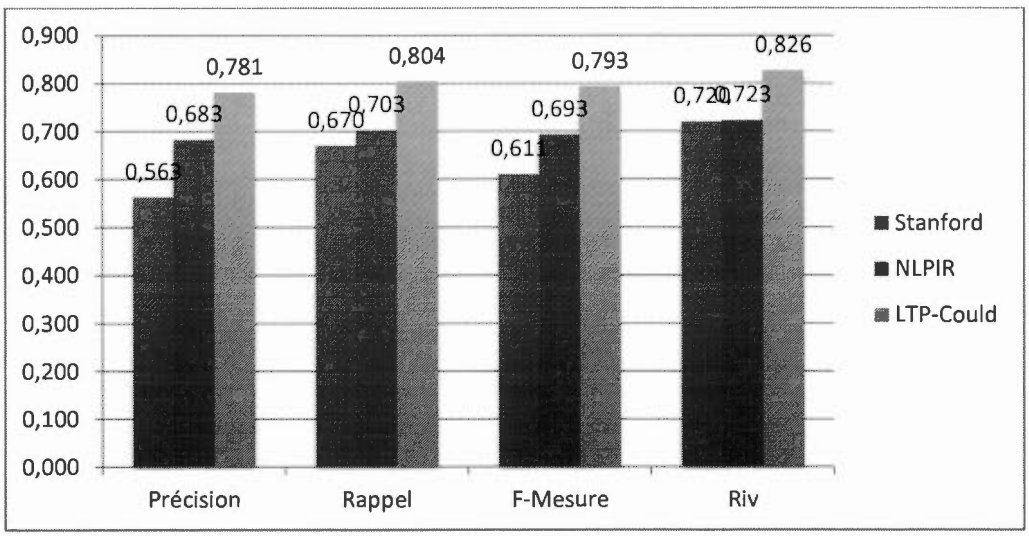
7. Les nouvelles façons d'utiliser des caractères : ce problème est dû à l'évolution de la langue chinoise. On rencontre à l'occasion, par exemple, 卧槽 (wo4cao2) sur les microblogues. Le premier caractère de cette séquence signifie « être couché » et le deuxième signifie « crèche ». Les deux caractères ensemble n'ont aucune signification, c'est-à-dire qu'ils ne forment pas un mot chinois. Mais selon la prononciation, et donc le pinyin des deux caractères, l'expression sans signification à l'écrit exprime un juron chinois (un « gros mot ») lorsqu'elle est prononcée oralement. En jouant ainsi sur la prononciation des caractères (et donc le pinyin), les utilisateurs peuvent utiliser des caractères complètement différents de ceux qu'ils devraient normalement utiliser.
8. Les mauvais caractères : si une personne a entré un mauvais caractère qui n'a pas le même pinyin que le hanzi exact, on ne peut pas le corriger.
9. La nouvelle forme apparue récemment « caractère (pinyin) caractère (pinyin) etc. » : nous avons présenté ci-dessus cette forme de problème qui est apparu depuis les dernières années dans la langue chinoise écrite.



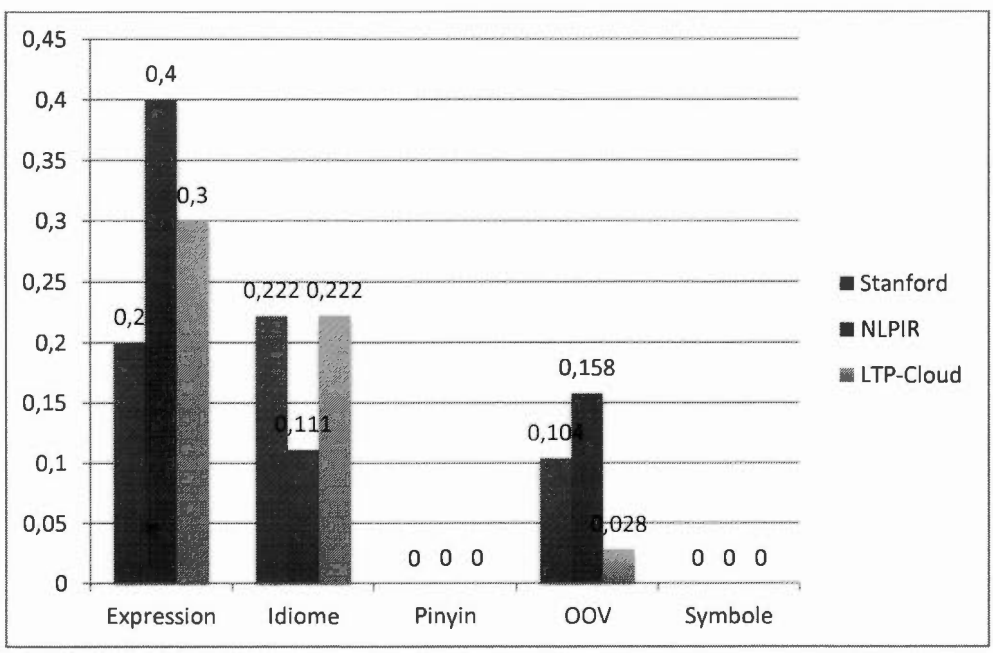
4.20 : Nombre de mots segmentés avec les trois outils Stanford, NLPIR et LTP-Cloud sur le petit corpus du microblogue

La figure 4.20 dévoile les faiblesses des trois outils quant au nombre de mots du texte original, car, entre autres, le texte publié dans les médias sociaux n'a pas une rigoureuse spécification syntaxique. Ce type de texte contient aussi beaucoup de symboles, d'OOV et d'autres nouvelles utilisations. Les figures 4.21 et 4.22 ci-dessous vont nous dévoiler la performance globale, ainsi que les statistiques des rappels pour les différents types de mots.

La performance des trois outils pour le texte publié dans les médias sociaux est plus faible que pour le texte général, un journal dans le cas étudié. Toutefois, nous notons que LTP-Cloud, à nouveau, offre une meilleure performance que les deux autres (figure 4.21). Nous pouvons trouver la performance des trois outils pour les différents types de mots.



4.21 : Statistique de la performance sur le petit corpus du microblogue



4.22 : Statistique des rappels de différents types de mots sur le petit corpus du microblogue

Comme nous le voyons à la figure 4.22, le LTP-Cloud offre une bonne performance relativement aux idiomes. La force du NLPIR réside quant à elle dans les expressions et les OOV. En revanche, la performance de Stanford n'est guère meilleure comparativement aux deux autres outils. Néanmoins, ces trois outils obtiennent 0 pour le pinyin et les symboles, c'est-à-dire qu'ils ne peuvent pas les segmenter, les prenant comme une partie du texte qui peut être ignorée. Notre recherche résout ce type de problème et, de ce fait, notre outil peut améliorer la performance de la segmentation pour ces types de mots.

4.10 Hypothèses et cheminement méthodologique proposé

Nous avons vu précédemment les différents problèmes relatifs à la segmentation du texte chinois, en particulier pour les textes tirés du microblogue Sina Weibo. À la section suivante, nous allons présenter premièrement notre proposition cognitive et informatique, puis proposer un système hybride comprenant sept modules. Ce système permet d'identifier le fonctionnement des symboles, la conversion du pinyin vers le hanzi et inversement, ce qui contribue nettement à l'optimisation de la performance de la segmentation des textes tirés des médias sociaux.

Ensuite, nous allons présenter le flux de données du système. Pour commencer, nous allons présenter le système dans son ensemble, c'est-à-dire la façon dont il peut résoudre les problèmes de segmentation du texte reliés au domaine général et celui que nous retrouvons dans les textes tirés de médias sociaux. Comme nous l'avons vu, ces textes présentent des problèmes spécifiques, comme ceux liés au pinyin ou aux symboles. Pour contourner ces difficultés, nous allons ajouter différents modules.

4.10.1 Volet cognitif

Dans son article, Wu (2011) soutient que la langue chinoise est davantage orientée sur la pragmatique que sur la syntaxe. Il résume cette thèse en trois points qui sont énumérés ci-dessous, puis propose un modèle cognitif pour pallier cette situation :

- Premièrement, la relation entre le sujet et l'objet sert à interpréter une situation plutôt que les caractéristiques des personnes ou des choses. Par exemple :

  1. 十个人 坐 一桌

     (Dix personnes) (s'asseoir) (une table)

  2. 一桌 坐 十个人

     (Une table) (s'asseoir) (dix personnes)

  Les deux phrases sont correctes, mais elles caractérisent deux situations différentes; la première signifie que nous avons invité dix personnes et chacune s'assoit à une table alors que la deuxième souligne la limite de la table : elle est pour dix personnes.

- Deuxièmement, on ne peut utiliser des règles purement syntaxiques pour spécifier le verbe entre le sujet et l'objet. Par exemple :

  1. 我们 今天 下 馆子

     (Nous) (Aujourd'hui) (En bas) (Restaurant)

     Aujourd'hui, nous allons chez un restaurant.

  2. 我们 打 的士

     (Nous) (Frapper) (Taxi)

     Nous prenons un taxi.

  Dans ce second exemple, on utilise 打 (da1, frapper) parce que c'est le mot qu'on utilise pour décrire les actions faites et que c'est qu'on appelle un taxi.

- Troisièmement, c'est le nom propre, comme le dévoilent aussi les travaux de Zheng et al. (2002). Ces mots n'ont pas un format standard en chinois et nous ne pouvons déterminer à l'avance combien de caractères ils contiennent.

Ces trois points, comme les erreurs que nous avons trouvées plus tôt, ont un impact important sur la performance de la segmentation. La figure 4.23 présente un modèle du processus humain de segmentation. Ce diagramme représente le travail cognitif d'une personne qui a une très bonne maitrise du chinois écrit.

Nous considérons qu'une phrase est entrée comme une séquence de caractères. Nous utilisons les numéros ci-dessous pour représenter les différents processus cognitifs exécutés par la personne. À chaque numéro correspond un processus dans le diagramme. Pour bien distinguer ce que la personne fait consciemment de l'action automatique de ses processus cognitifs, nous avons mis en gras les processus de niveau conscients (Figure 4.23).

Lorsqu'une personne lit du texte chinois :

1. Un hanzi est entré.
2. La signification de ce(s) caractère(s) est vérifiée selon un dictionnaire interne.
3. Si ce caractère correspond à un mot, ce mot est temporairement mémorisé. Sinon, le processus automatique de lecture retourne à l'étape 1 et entre un second caractère.
4. Le caractère suivant de ce mot temporaire est entré.
5. La relation entre ce nouveau caractère et le mot temporairement mémorisé est vérifiée selon un dictionnaire interne.
6. Si le mot temporairement mémorisé additionné du caractère suivant peut devenir un nouveau mot, alors ce nouveau mot va prendre la place du mot temporairement mémorisé, et le processus revient à l'étape 4. Sinon, une segmentation va être faite à cet endroit: le processus automatique de lecture va considérer ce qui précède comme un mot (**c'est-à-dire que la personne a**

perçu un mot), et revenir à l'étape 1 du processus **(c'est-à-dire que la personne poursuit sa lecture)**.

Les étapes 2 et 5 considèrent la *signification* des mots. En particulier, nous pouvons mettre à jour le dictionnaire en y ajoutant des mots du domaine d'expertise, des mots spécifiques, des symboles, etc.

Ici, le diagramme nous montre comment une personne fait la segmentation pendant la lecture. L'article de Li et al. (2009) se concentre sur l'influence de la reconnaissance du premier mot par celle des caractères suivants ou du mot suivant pendant la lecture. Ils ont proposé un modèle simple de la segmentation et de la reconnaissance de mots basé sur une séquence de quatre hanzi. Ils ont considéré cette séquence selon cinq différentes conditions :

1) un seul mot de quatre caractères ;

2) deux mots de deux caractères, où il n'existe pas de relation sémantique entre ces deux mots ;

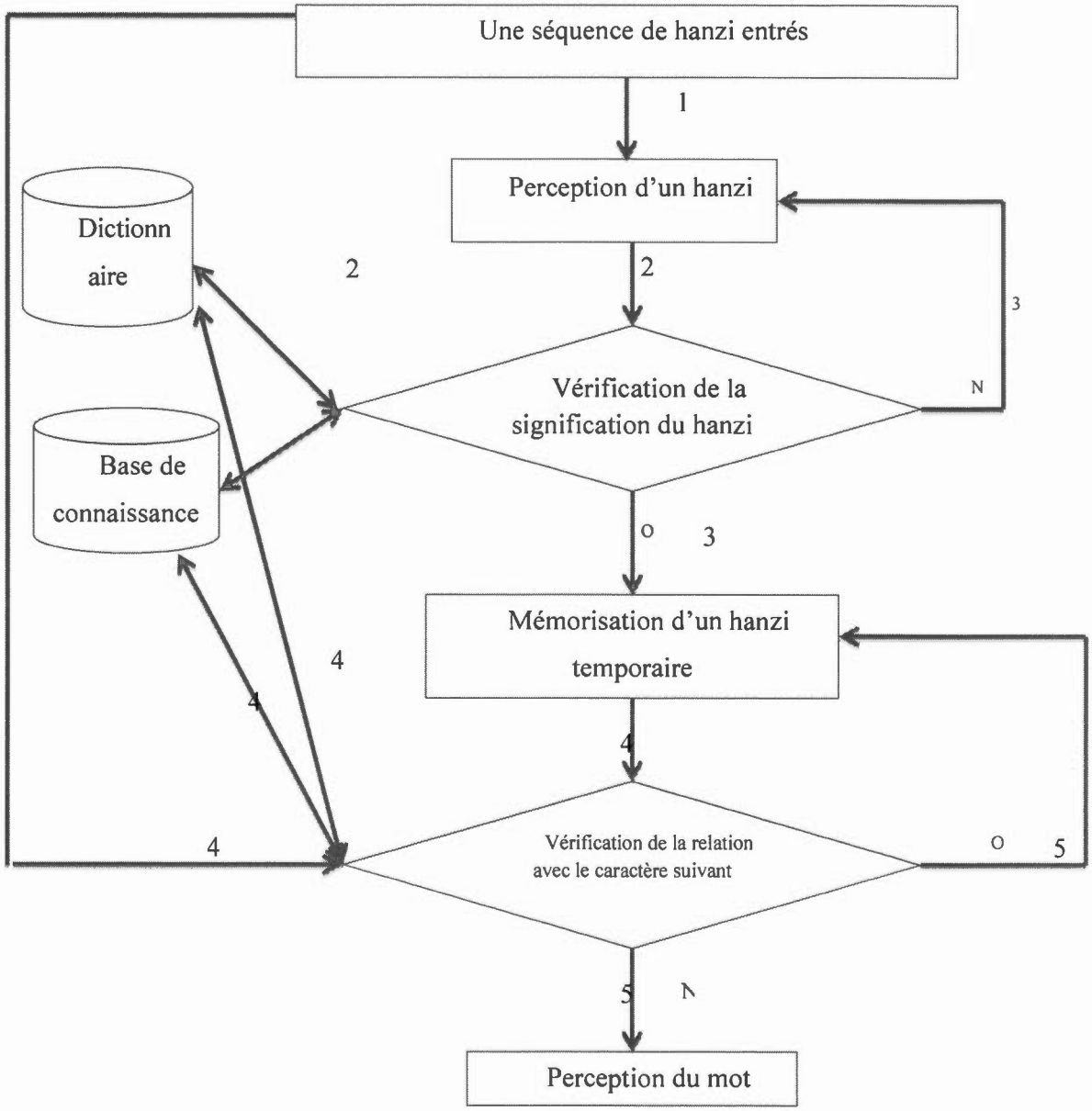
3) un mot de deux caractères avec les deux caractères individuels suivants ;

4) quatre caractères individuels ;

5) deux mots de deux caractères, où il existe de relation entre ces deux mots.

Les auteurs ont ensuite analysé l'influence de la reconnaissance du premier mot sur celle des autres (par le biais de cinq expérimentations), et ont trouvé que le taux de précision des conditions 1 et 3 est le plus élevé ; ensuite viennent la condition 2 puis les conditions 4 et 5, qui sont les plus basses. Ils ont ainsi montré que la reconnaissance des mots ou caractères précédents est influencée par celle des mots suivants si ces derniers ont une relation sémantique avec le premier. Si le premier mot n'a aucune relation sémantique avec les caractères ou le mot suivant, alors il est plus facile de segmenter ou reconnaître ce mot. Le problème de la segmentation des textes et celui de la reconnaissance d'un mot sont comme le paradoxe de l'œuf et de la poule : la reconnaissance d'un mot est basée sur le texte segmenté, et la segmentation des textes est basée sur la reconnaissance des mots de la langue chinoise. Ce modèle a

aussi montré les limites de la segmentation quand le troisième caractère d'une séquence de quatre caractères peut se combiner avec le mot précédent pour constituer un mot ou avec le caractère suivant pour constituer un autre mot, et où il faut de l'information sémantique du contexte pour déterminer laquelle est la bonne segmentation.



4.23 : Processus cognitif de la segmentation du texte chinois

4.10.2 Volet informatique

Sur ce plan, notre objectif est de construire un système informatique qui contient les différents modules cognitifs décrits ci-dessus pour réaliser le processus humain entier de lecture du texte chinois, incluant son étape fondamentale de segmentation. Pour étendre l'usage de ce système informatique à divers utilisateurs ou dans divers contextes, nous voulons développer en particulier un système à partir duquel nous pouvons ajouter, modifier ou cacher les modules en fonction des différents types de textes. Les différentes fonctions des modules informatiques de notre système sont expliquées ici.

- ***Module du dictionnaire.*** Ce module contient les mots généraux qui sont nécessaires à la compréhension du chinois de base. Nous pouvons aussi y ajouter les mots spécifiques de divers domaines (médical, technique, etc.). À la fin de chaque segmentation, le système peut aussi mettre à jour ce dictionnaire en y ajoutant les nouveaux mots découverts lors du processus de segmentation. Une faible probabilité sera d'abord accordée aux nouveaux mots, laquelle sera renforcée à chaque fois qu'il est « découvert » par le système.
- ***Module de la base de connaissance des idiomes.*** Ce module développé par Wang et al. (2013) contient les idiomes généraux. Il peut lui aussi se mettre à jour à la fin de la segmentation en ajoutant les nouveaux idiomes découverts lors du processus. Une autre fonction de ce module consiste à identifier avec précision les changements apportés dans les idiomes existants, par exemple, les hanzi qui ont été remplacés et retrouvés dans les idiomes originaux.
- ***Module du pinyin.*** Ce module est spécifique pour les textes de microblogues. Il peut identifier les pinyins dans le texte et les distinguer de l'anglais ou du français. Ce module contient aussi un outil de segmentation du pinyin. Ainsi, quand il rencontre une séquence de pinyin, il peut facilement l'identifier, car il contient également un dictionnaire répertoriant le pinyin et les hanzi

correspondants. Lorsque des entrées incorrectes de hanzi sont faites, il peut immédiatement vérifier l'orthographe et procéder à nouveau à la conversion du pinyin aux hanzi. Comme pour les modules précédents, il peut aussi être mis à jour à la fin du processus.

- ***Module des symboles.*** Ce module est lui aussi spécifique aux textes de microblogues. Il contient un dictionnaire qui peut identifier les symboles et leurs fonctions dans le texte.
- ***Module de vérification syntaxique.*** Ce module est utilisé pour vérifier la syntaxe ou la structure des phrases (les règles).
- ***Module de calcul des probabilités.*** Ce module utilise le réseau bayésien pour combiner la probabilité des mots, la probabilité entre les caractères et la probabilité des facteurs d'impact (Importance et Indice). À la fin du processus, il peut augmenter ou diminuer la probabilité des mots dans les dictionnaires et bases de connaissances, ce qui ne peut qu'améliorer la performance lors de la prochaine segmentation.

Tous les nouveaux modules sont programmés en Java. Nous allons également fournir une interface utilisateur qui permet d'entrer le texte à segmenter et, si nécessaire, de télécharger son propre dictionnaire spécialisé vers le système. Nous allons considérer tous les facteurs influençant la segmentation et utiliser le réseau bayésien pour calculer les probabilités. Au terme de ces étapes, le système va nous donner la probabilité de satisfaction de la segmentation d'une phrase. Nous jugerons avoir atteint la fin du processus si, après des plusieurs exécutions, cette probabilité atteint une valeur stable, c'est-à-dire que la segmentation sera devenue stable.

4.10.3 Hypothèses et cheminement méthodologique proposé

Voici le diagramme du système intégral que nous proposons, incluant nos propres modules tels que décrits à la section précédente, ainsi que les modules provenant du domaine public (notamment de l'université Stanford). Dans ce système, nous avons tenté de résorber tous les problèmes rencontrés lors de l'analyse des erreurs de segmentation.

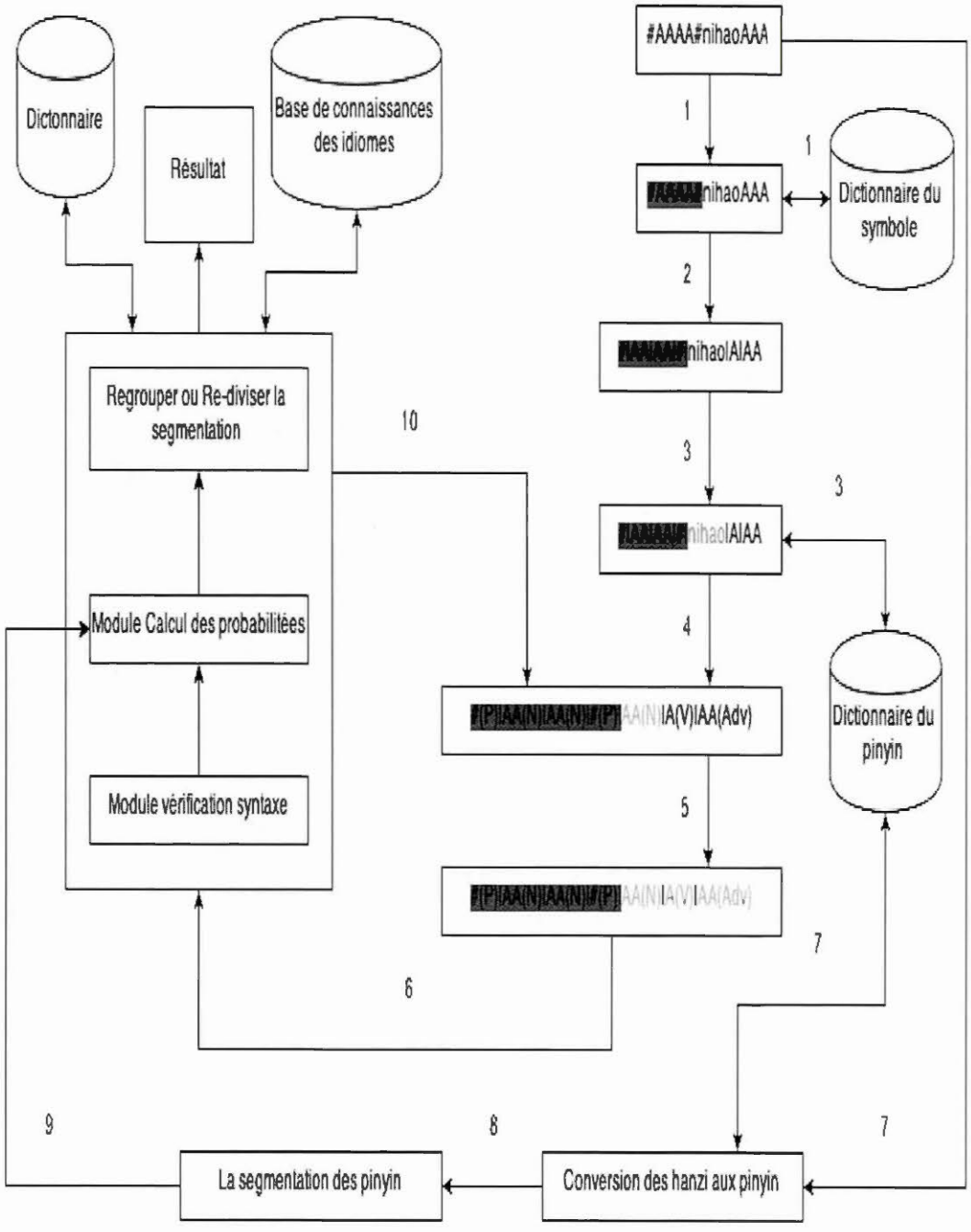
La figure 4.24 ci-dessous représente le système en entier. Quant à la figure 4.25, elle donne un exemple du flux de données du système. Notre système est basé sur trois outils développés par l'université de Stanford:

- ***Le segmenteur de Stanford (Stanford Segmenter).*** Nous utilisons le segmenteur de Stanford pour initialiser la segmentation. Celui-ci fournit donc un texte segmenté pour l'étape suivante.
- ***Le système de POS tagging de Stanford (Stanford POS tagging).*** Nous utilisons le système de POS tagging de Stanford pour assigner une étiquette POS (Part-Of-Speech) à chaque mot, c'est-à-dire pour associer aux mots du texte des informations grammaticales correspondantes (comme le rôle syntaxique, le genre, le nombre, etc.), laquelle pourra aider l'analyse grammaticale et syntaxique, comme dans les travaux effectués par Qian et Liu (2012).
- ***L'analyseur syntaxique de Stanford (Stanford Parser).*** Nous utilisons l'analyseur syntaxique de Stanford pour construire un arbre grammatical de la phrase, comme dans les recherches de Li et al. (2010). Cela nous permet ensuite de labelliser les erreurs grammaticales et syntaxiques (par un système à base de règles).

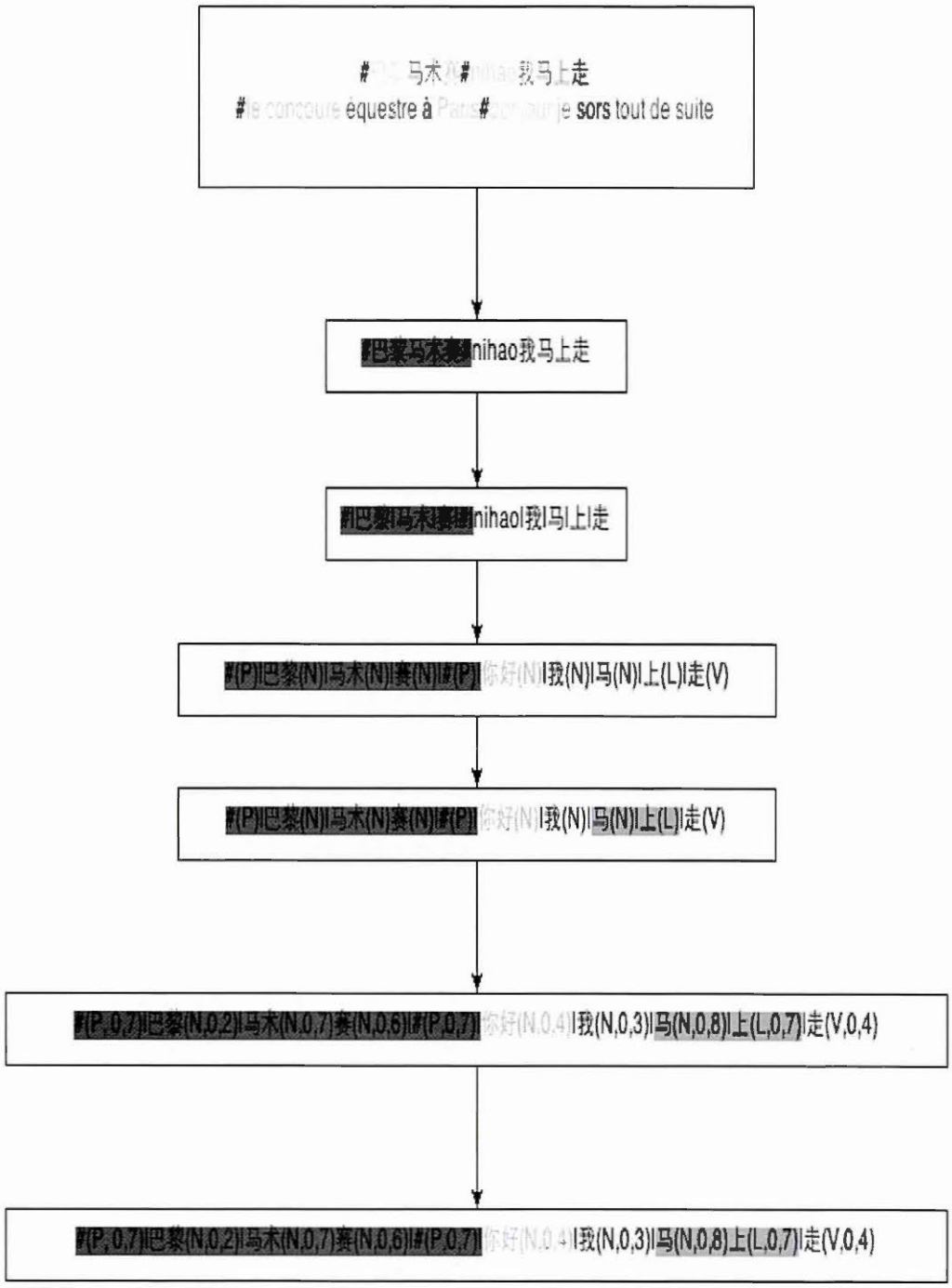
Le flux des données est représenté à la figure 4.25, dans laquelle les numéros correspondent aux fonctions énumérées :

1. Vérifier et labelliser la fonction des symboles avec un dictionnaire des symboles, en particulier pour les textes de microblogues.
2. Segmenter le texte chinois par le segmenteur de Stanford.
3. Marquer la présence de pinyin au moyen d'une étiquette (par exemple /P ni2hao3 /P) en utilisant un dictionnaire pinyin.
4. Étiqueter les mots par le système de POS tagging de Stanford.
5. Utiliser l'analyseur syntaxique de Stanford pour l'analyser la syntaxe.
6. Entrer le résultat dans le module de vérification de la syntaxe.
7. Convertir les hanzi du texte en pinyin avec le dictionnaire pinyin.
8. Segmenter la séquence de pinyin.
9. Entrer les résultats dans le module de calcul des probabilités. Nous allons présenter ce module dans la section suivante. Ce module ajuste les probabilités des mots, pour ensuite les regrouper ou les rediviser en fonction des probabilités calculées.
10. Les nouvelles segmentations de mots retournent à l'étape 4.

Ce processus se termine lorsque les probabilités calculées par le module de calcul des probabilités deviennent stables.



4.24 : Système cognitif proposé de la segmentation du texte chinois



4.25 : Exemple de segmentation

Les fonctions des différents modules étant développées dans le paragraphe précédent, nous allons maintenant donner un exemple pour présenter le fonctionnement du processus du système.

- Une phrase entre dans le système.
- La phrase passe ensuite dans le module d'analyse des symboles. Ce module peut identifier les fonctions des symboles dans le texte. Par exemple, le système identifiera #巴黎马术比赛# (#ba1li2ma3shu4bi3sai4#, le concours équestre à Paris) comme un sujet et @天空的云 (@tian1kong1deyun2, le nuage au ciel) comme le nom d'une personne.
- Le résultat passe ensuite au module du pinyin. S'il existe une séquence de lettres correspondant à du pinyin dans la phrase, ce module peut l'identifier et la segmenter. Ce module va également convertir tous les mots à leur pinyin pour le module de calcul des probabilités. Si, selon le dictionnaire pinyin, ces lettres ne correspondent à aucun pinyin, le module va les analyser comme un ensemble de mots formés de lettres alphabétiques (français, anglais, etc.).
- Le segmenteur de Stanford segmente la phrase. Puisque le texte a déjà été labellisé par le module d'analyse des symboles, nous avons donc besoin de remettre les caractères ensemble, parce que le segmenteur Stanford ne comprend pas les fonctions des symboles. Il proposera, par exemple, le texte segmenté suivant : #|巴黎|马术|比赛|#

  Nous pouvons considérer que cette expression entière est un nom spécifique et la mettre à jour dans le dictionnaire. Si cette expression apparaît dans une autre phrase ne possédant pas le même format (c'est-à-dire, pas entre des #), alors le segmenteur fera la segmentation entre ces mots. Selon l'aspect de la compréhension, nous remettons ces mots ensemble. Cette expression entière peut aider à augmenter la performance de la traduction automatique.

- Le module de POS tagging de Stanford assigne une étiquette POS à chaque mot.
- Le module d'analyse syntaxique de Stanford construit l'arbre grammatical de la phrase.
- Le module de vérification de la syntaxe vérifie la syntaxe de la phrase. Par exemple, quand il voit une préposition avec un adverbe, il fait une marque d'erreur parce que, selon la grammaire du texte chinois, une préposition vient seulement avec un nom ou un pronom qui présente un lieu, une heure, un statut, etc.
- La phrase avec ses étiquettes symboliques POS et ses marques syntaxiques entre dans le module de calcul des probabilités. D'abord, ce module va calculer la fréquence des mots, les probabilités en 5-gram entre les mots et le calcul identique entre le pinyin. Ensuite, il va considérer tous les facteurs pour ajuster les probabilités entre les mots. Il existe plusieurs situations motivant un ajustement.
    1. Si, après la comparaison avec la probabilité entre deux pinyins, le système juge qu'il existe de mauvaises entrées de hanzi correspondant à des mots en pinyin, le module va augmenter la probabilité entre ces mauvaises entrées. Si, par exemple, le système reçoit les deux caractères 尼好, le segmenteur séparera ces deux caractères, car il existe une très basse probabilité entre eux (ils ne forment pas un mot chinois). Cependant, selon leur pinyin, ces caractères correspondent au mot « bonjour » en chinois, où le premier caractère est une mauvaise entrée (ayant le même pinyin avec le bon caractère), et donc le module augmentera la probabilité entre ces deux caractères.
    2. Si un mot ou des mots ont une étiquette syntaxique, le module va séparer ce(s) mot(s) en caractères et calculer la probabilité en 5-gram entre ces caractères et entre les caractères antérieurs et postérieurs. Le module

augmentera la probabilité de ce(s) mot(s) et considèrera ces nouvelles probabilités en 5-gram, c'est-à-dire qu'il augmentera les probabilités entre les caractères parce que ceux dont la probabilité de cooccurrence était avant très basse sont apparus ensemble comme un mot dans une phrase.

3. Selon la base de connaissance des idiomes, le module augmentera les probabilités entre les sous-parties de l'idiome. Par exemple, si le segmenteur divise un idiome en deux parties, le module pourra augmenter la probabilité entre ces parties si, après la comparaison avec la base de connaissance des idiomes, il trouve qu'elles forment un idiome.
4. S'il y a encore des problèmes qui n'ont pas été identifiés, nous pourrons en tenir compte dans des versions futures du projet.

- Après l'ajustement des probabilités, le module de regroupement des mots va faire son travail selon leurs nouvelles probabilités et mettre à jour les nouveaux mots, les nouvelles expressions et les nouveaux idiomes dans un dictionnaire.
- La phrase doit ensuite repasser par le module du segmenteur de Stanford; le processus reproduit la démarche jusqu'à ce que les probabilités deviennent stables.

CHAPITRE V

ÉVALUATION ET DISCUSSION

Les résultats obtenus par notre système de la segmentation sont présentés et expliqués dans ce chapitre. Nous testons celui-ci en utilisant le même double corpus, soit un corpus provenant d'un journal et un autre provenant des médias sociaux. Ensuite, nous comparons notre résultat avec celui de trois autres outils de segmentation. Pour chaque comparaison, nous donnons également des exemples de fonctionnement pour expliquer la raison pour laquelle la performance de notre système est supérieure à celle des autres segmenteurs.

5.1 Résultats obtenus sur le petit corpus du journal

		Mot	Expression	Idiome	OOV
Segmenteur Stanford	Nombre	1 463	7	4	2
	Précision	0,714	-	-	-
	Rappel	0,831	0,185	0,214	0

	F-mesure	0,768	-	-	-
	Riv	0,856	-	-	-
NLPIR (ICTCLAS)	Nombre	1 245	10	5	4
	Précision	0,793	-	-	-
	Rappel	0,786	0,296	0,357	0,028
	F-mesure	0,789	-	-	-
	Riv	0,809	-	-	-
LTP-Cloud	Nombre	1 197	13	4	5
	Précision	0,831	-	-	-
	Rappel	0,792	0,407	0,286	0,028
	F-mesure	0,811	-	-	-
	Riv	0,816	-	-	-
Notre Système	Nombre	1 228	15	9	24
	Précision	0,873	-	-	-
	Rappel	0,854	0,481	0,643	0,583
	F-mesure	0,864	-	-	-
	Riv	0,862	-	-	-

5.1 : Comparaison entre les systèmes sur le petit corpus du journal

Au tableau 5.1, nous constatons que notre système segmente le corpus en 1 228 mots (voir la figure 5.2), dont 15 expressions, 9 idiomes et 24 OOV, corpus,

avec comme résultat, 1 073 mots, 13 expressions, 9 idiomes et 21 OOV analysés avec exactitude. Le nombre d'idiomes augmente pour deux raisons. La première est qu'en raison du dictionnaire des idiomes que nous utilisons, le système peut détecter les idiomes dans le corpus. La deuxième raison est liée à la segmentation du pinyin : selon la 8e étape du flux des données exposé ci-dessus, lorsque la séquence de pinyin est analysée par l'outil de segmentation du pinyin, elle se transforme en unités lexicales identifiées par le dictionnaire pinyin. L'exemple ci-dessous provenant du corpus peut expliquer la façon dont notre système se montre capable de trouver un nouvel idiome, qui est une variante de l'idiome originel. Cet exemple révèle que le fonctionnement de notre système permet de résoudre la deuxième difficulté identifiée dans la section 2.4 de la problématique.

Prenons l'exemple d'une séquence de l'idiome variant[46] en pinyin comme : gua1mu4xiang1kan4 (瓜目相看). Le segmenteur de Stanford segmente la séquence de hanzi comme 瓜|目|相看 (gua1, le melon ; mu4, l'œil; xiang1kan4, se regarder). Dans cet exemple, la personne utilise le caractère 瓜 (gua1, le melon) pour remplacer le caractère 刮 (gua1, gratter). L'idiome originel est 刮目相看 (gua1mu4xiang1kan4, avoir une appréciation toute nouvelle sur quelqu'un d'impressionnant). Les deux caractères ont le même pinyin et le même ton. Un système qui utilise uniquement les outils de la segmentation hanzi ne peut pas reconnaitre cet idiome. Toutefois, lorsque notre système convertit la phrase en hanzi au pinyin, le segmenteur du pinyin peut facilement la reconnaitre grâce au dictionnaire pinyin. Dans ce cas, le noyau du système, le réseau bayésien, va augmenter la probabilité entre le pinyin gua1 et le pinyin mu4, et le pinyin mu4 et le pinyin xiang1kan4.

La segmentation du pinyin joue le même rôle dans l'amélioration de la performance de la segmentation des expressions, comme le montre l'exemple suivant.

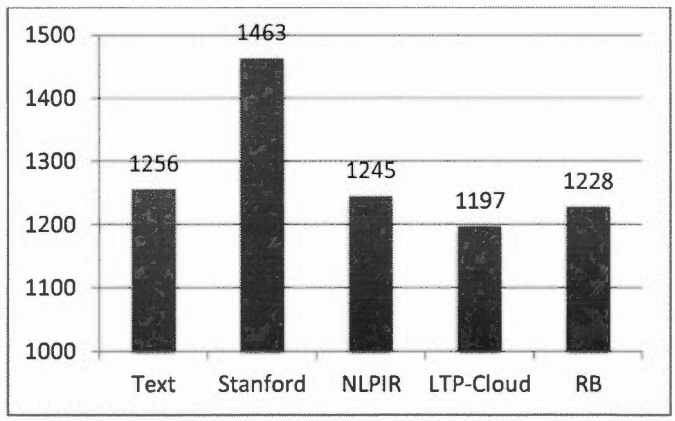
---

[46] Ici, le(s) caractère(s) des idiomes est (sont) remplacé(s) par les autres caractères; ce nouvel idiome représente la signification originelle ou une nouvelle signification.

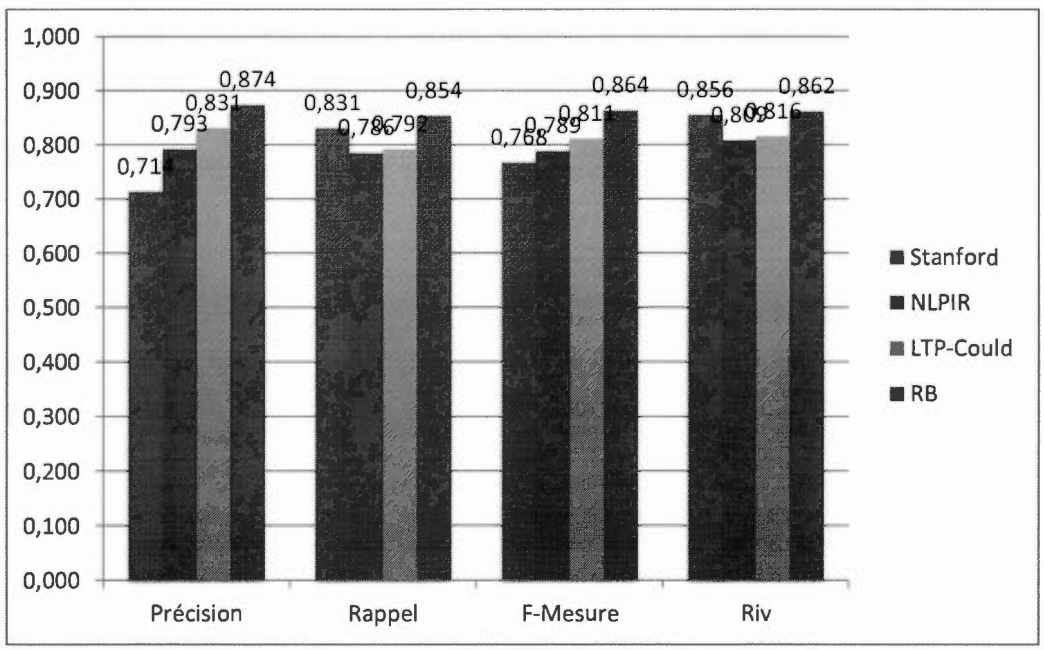
Prenons une séquence de pinyin comme : shan1qian2liu2ming2 (山前刘明). Le segmenteur de Stanford segmente l'expression 山前刘明 comme 山前|刘明 (shan1qian2, avant la montagne; liu2ming2, un nom de la personne). La séquence ne signifie pas qu'une personne est avant la montagne. Celui qui a écrit cette expression utilise le mot 山前 (shan1qian2, avant la montagne) pour remplacer le mot 删前 (shan1qian2, avant supprimer), et le mot 刘明 (liu2ming2, un nom de la personne) pour remplacer le mot 留名 (liu2ming2, laisser le nom), cette expression originelle est shan1qian2liu2ming2 (删前留名, laisser le nom avant de supprimer les commentaires). Par conséquent, en fonction du pinyin segmenté, le réseau bayésien va augmenter la probabilité entre les transcriptions pinyin, shan1qian2 et liu2ming2. Cette fonction permet de pallier à la première difficulté soulevée à la section 3.3 (Problématique).

Les deux exemples précédents confirment l'avantage de notre système, qui utilise un segmenteur du pinyin pour trouver les bons hanzi associés. Le corpus du journal possède une grammaire standard et une syntaxe stricte, aussi le dictionnaire de symboles et la vérification grammaticale ne jouent qu'un rôle minime dans l'évaluation de performance.

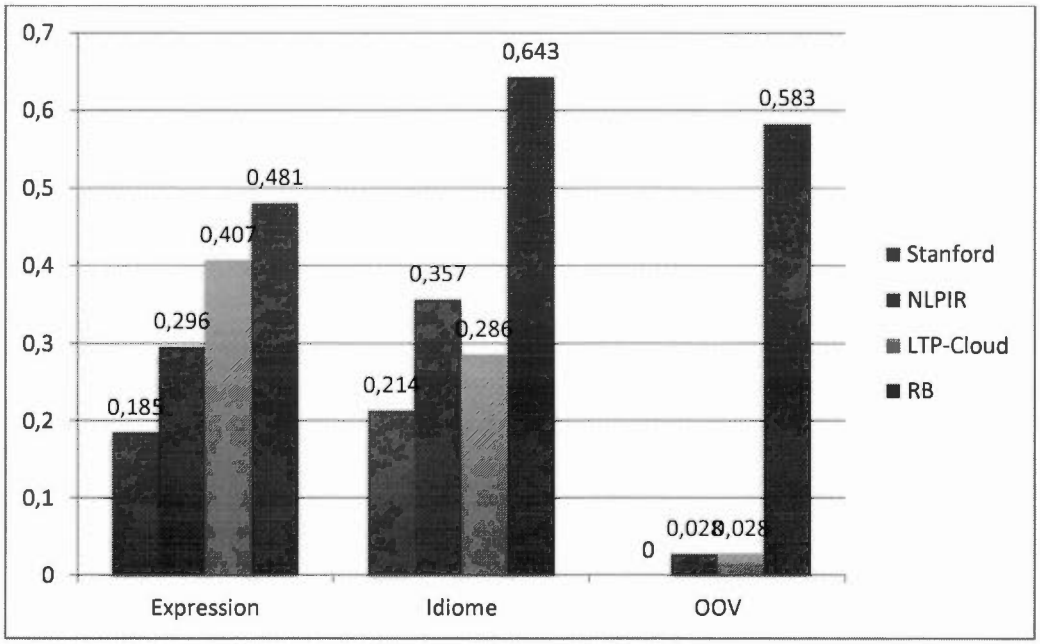
La figure 5.3 rapporte que notre système présente une haute valeur du F-mesure en combinant les deux segmentations. Nous constatons également à la figure 5.4 que notre système fait preuve d'une haute valeur de rappel, c'est-à-dire que le taux correct du nombre d'expressions, d'idiomes et d'OOV dévoilé par notre système est plus élevé que pour les trois autres systèmes.



5.2 : Nombre des mots sur le petit corpus du journal



5.3 : Statistique de la performance sur le petit corpus du journal



5.4 : Statistique des rappels de différents tris de mots sur le grand corpus du journal

5.2 Résultats obtenus sur le petit corpus des médias sociaux

		Mot	Expression	Idiome	Pinyin	OOV	Symbole

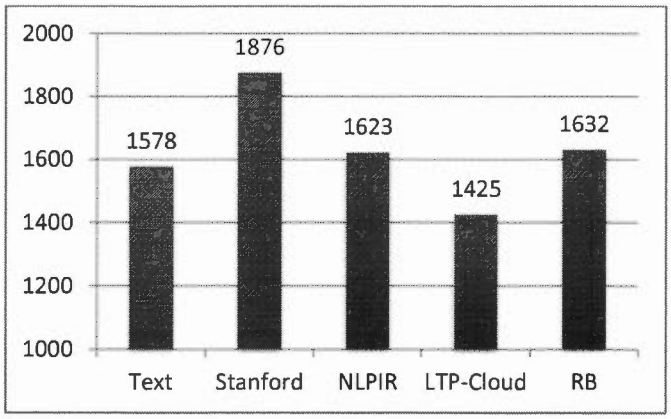
Corpus Médias Sociaux

9	-	0	-	-	11	-	0	-
14	-	0,104	-	-	17	-	0,158	-
3	-	0	-	-	3	-	0	-
5	-	0,222	-	-	4	-	0,111	-
5	-	0,200	-	-	8	-	0,400	-
1 876	0,563	0,670	0,611	0,720	1623	0,683	0,703	0,693
Nombre	Précision	Rappel	F-mesure	Riv	Nombre	Précision	Rappel	F-mesure
Segmenteur Stanford					NLPIR (ICTCLAS)			

-	11	-	0	-	-	24	-	0,479		
-	21	-	0,211	-	-	37	-	0,491		
-	3	-	0	-	-	9	-	0,449		
-	6	-	0,222	-	-	6	-	0,465		
-	7	-	0,300	-	-	7	-	0,510		
0,723	1 425	0,781	0,804	0,793	0,826	1 632	0,892	0,923		
Riv	Nombre	Précision	Rappel	F-mesure	Riv	Nombre	Précision	Rappel		
L'TP-Cloud					Notre système					

	F-mesure	0,907	-	-	-	-	-
	Riv	0,939	-	-	-	-	-

5.5 : Comparaison entre les systèmes pour le petit corpus des médias sociaux



5.6 : Nombre de mots segmentés avec les quatre outils Stanford, NLPIR, LTP-Cloud et RB sur le petit corpus des médias sociaux

Nous utilisons un corpus qui contient au total 3256 hanzi[47]. Au tableau 11, nous constatons que notre système a trouvé 1 632 mots, dont 7 expressions, 6 idiomes, 9 pinyin, 37 OOV et 24 symboles. Dans les figures 32 et 33, il semble évident que les

[47] Ce corpus vient du site www.nlpir.org, et le période du micro-blog SINA du 12 Octobre 2009 au 11 Novembre 2009.

valeurs de F-mesure et de rappel sont plus hautes que les autres trois systèmes, les raisons en sont les suivantes :

1. Les trois systèmes n'utilisent pas le segmenteur du pinyin et n'ont pas intégré un dictionnaire pinyin. Par ailleurs, quand le système rencontre une séquence de pinyin, il la traite toujours comme des mots en anglais.

Par exemple :

La séquence de pinyin : nihaoma

Les trois systèmes détectent la séquence comme un mot anglais alors que ce n'est pas le cas. Lorsque notre système analyse cette séquence avec la segmentation hanzi, il la détecte également comme un mot anglais. Or, avec la segmentation du pinyin, le segmenteur la segmente correctement : nihao|ma. Selon le dictionnaire du pinyin, elle correspond aux hanzi 你好|吗 (ni3hao3, bonjour ; ma1, un mot particule). La probabilité entre les deux mots augmente, ce qui rend la performance de la segmentation finale plus efficace.

2. Comme nous l'avons présenté dans la section 5.1, notre système de la segmentation peut détecter les variantes d'idiome et d'expression en utilisant le segmenteur du pinyin.
3. Le dictionnaire du symbole contribue également à identifier les symboles dans le corpus des médias sociaux.

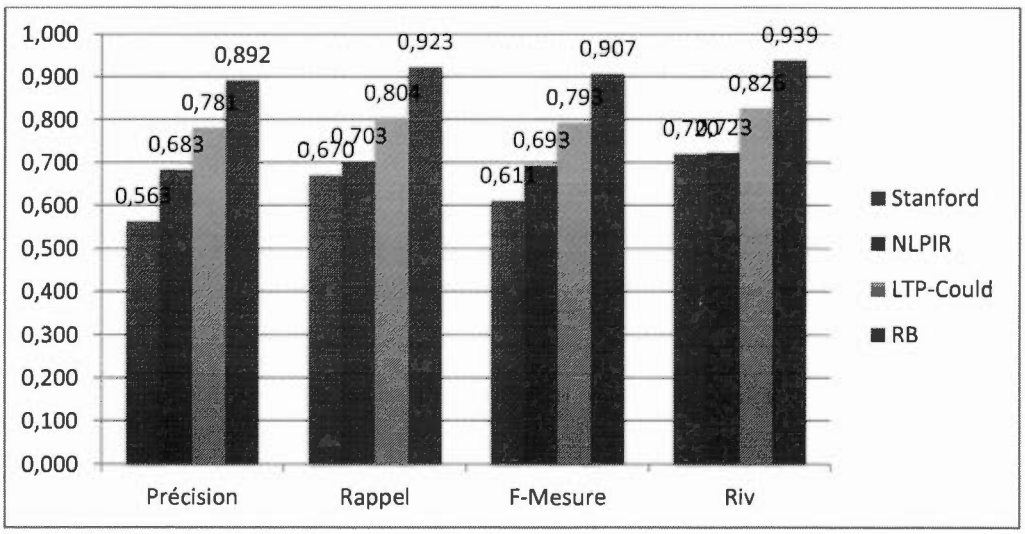
Par exemple :

88 – 拜拜 (bai1bai1, au revoir)

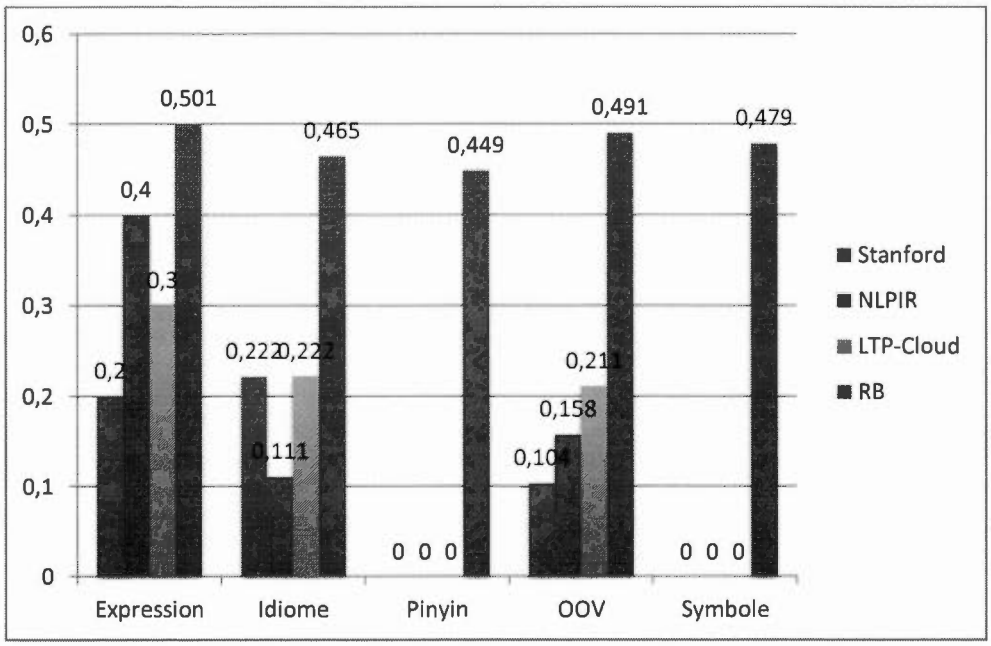
4. Le *Semantic Chunk* peut aussi augmenter la performance de la segmentation. Le corpus tiré des médias sociaux que nous utilisons est un extrait de Weibo pris en ordre chronologique avec, par conséquent, une succession de commentaires en général autour d'un thème. Pour calculer les fréquences des mots, en excluant par exemple les mots d'arrêt et les mots fonctionnels, nous

divisons le grand corpus en différents *chunks*, chaque *chunk* autour d'un thème, d'un mot ou de plusieurs mots plus fréquemment employés. Si un ensemble de caractères absent du dictionnaire est fréquent dans un *chunk*, nous pouvons en conclure que cet ensemble de caractère est un mot OOV et peut-être un nouveau mot. C'est un aspect du *Semantic Chunk* qui permet d'améliorer la performance de la segmentation. Un autre aspect du *Semantic Chunk* réside dans le fait que la probabilité entre les mots est ajustée selon la fréquence des mots. Par exemple, si un mot OOV est plus fréquent dans un *chunk*, les probabilités entre les caractères dans ce mot OOV vont augmenter et donc influencer le réseau bayésien.

5. La fonction qui traite la forme de la nouvelle structure contribue aussi à amplifier la performance de la segmentation. Tel que mentionné précédemment, la nouvelle structure hanzi (pinyin) hanzi (pinyin) se retrouve dans les textes des médias sociaux. Grâce au dictionnaire pinyin, le système identifie les deux mots (hanzi et hanzi, pinyin et pinyin) au lieu de quatre caractères.



5.7 : Statistique de la performance du petit corpus des médias sociaux



5.8 : Statistique des rappels de différents types de mots sur le petit corpus des médias sociaux

5.3 Résultats obtenus sur un grand corpus du journal

Dans le texte précédent, nous avons utilisé un corpus qui contient au total 2 216 hanzi et 1 256 mots. Notre système montre un meilleur résultat que les autres trois segmenteurs. Ensuite, nous utilisons un corpus qui contient au total 23 106 hanzi[48]. Le tableau 5.10 montre les statiques de la segmentation effectuée par les trois outils et notre système pour le texte général.

[48] Ce corpus vient du site www.nlpir.org, et le période du journal du 12 Octobre 2009 au 14 Décembre 2009.

	Hanzi	Mot	Expression	Idiome	Pinyin	OOV	Symbole
Nombre	23 106	15 352	2 357	386	0	759	0
% par rapport le mot	-	-	15,35%	2,51%	0%	4,94%	0%

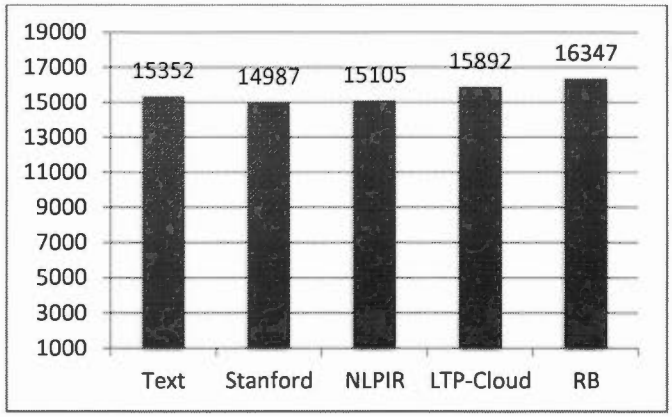
5.9 : Statistique du nombre de tailles du grand corpus

		Mot	Expression	Idiome	OOV
Segmenteur Stanford	Nombre	14 987	2 175	296	563
	Précision	0,815	-	-	-
	Rappel	0,796	0,704	0,624	0,149
	F-mesure	0,805	-	-	-
	Riv	0,829	-	-	-
NLPIR (ICTCLAS)	Nombre	15 105	2 486	374	956
	Précision	0,765	-	-	-
	Rappel	0,753	0,903	0,866	0,335
	F-mesure	0,759	-	-	-
	Riv	0,774	-	-	-
LTP-Cloud	Nombre	15 892	2 698	348	898

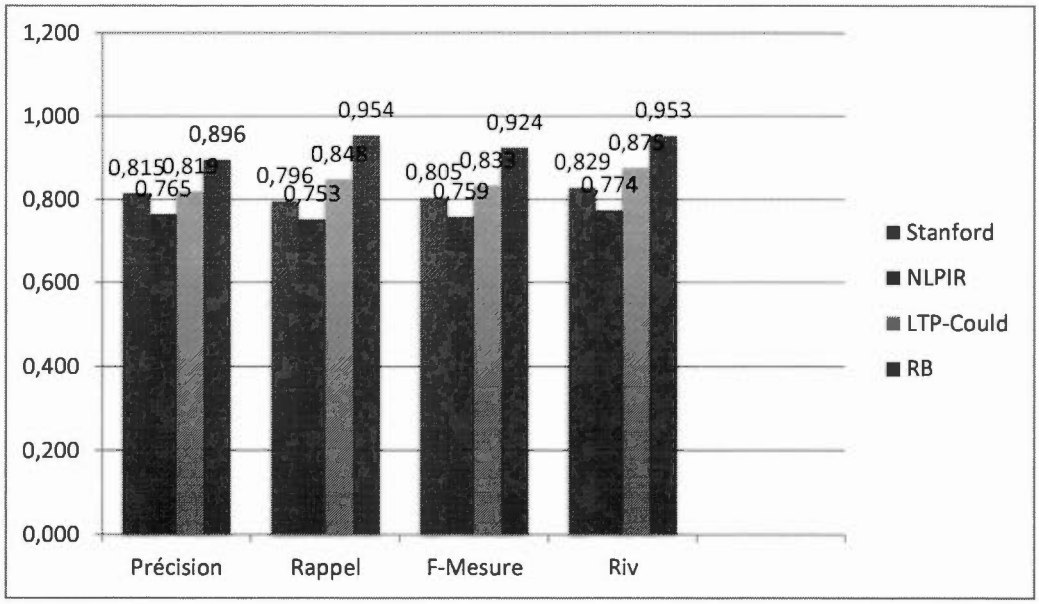
	Précision	0,819	-	-	-
	Rappel	0,848	0,976	0,816	0,325
	F-mesure	0,833	-	-	-
	Riv	0,875	-	-	-
Notre Système	Nombre	16 347	2 579	365	864
	Précision	0,896	-	-	-
	Rappel	0,954	0,960	0,912	0,973
	F-mesure	0,924	-	-	-
	Riv	0,953	-	-	-

5.10 : Statistique de la segmentation effectuée par les trois outils avec notre système pour le texte général

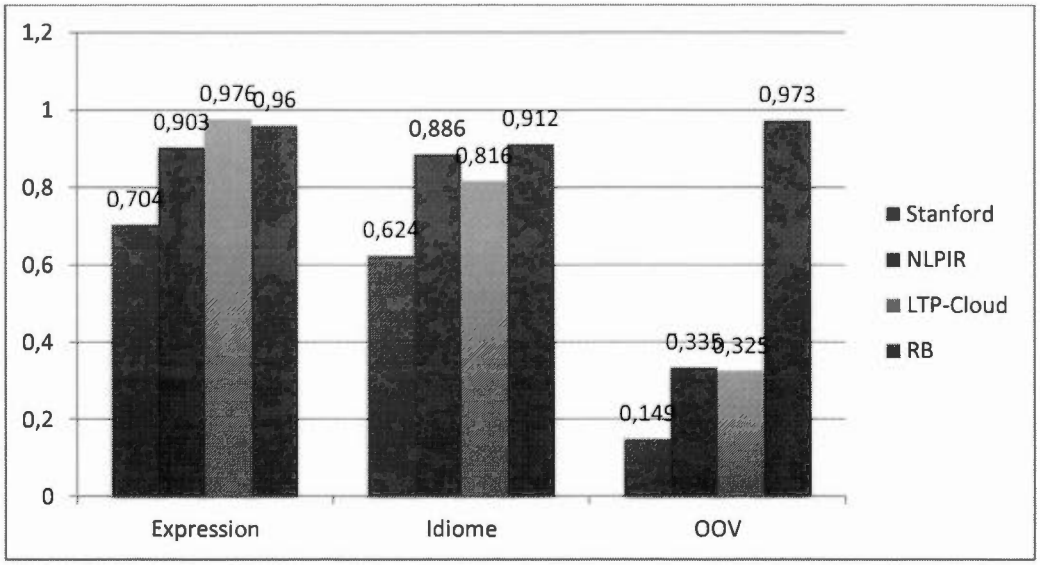
Comme nous le voyons au tableau 5.11, le segmenteur Stanford considère que le corpus contient 14 987 mots, NLPIR qu'il y en a 15 105, LTP-Cloud 15 892 mots et notre système 16 347 mots.



5.11 : Nombre de mots du grand corpus du journal



5.12 : Statistique de la performance de la segmentation du grand corpus du journal



5.13 : Statistique des rappels de différents types de mots sur le grand corpus du journal

La figure 5.12 rapporte que notre système présente une haute valeur de F-mesure en combinant les deux segmentations. Nous constatons également à la figure 5.13 que notre système fait preuve d'une haute valeur de rappel, c'est-à-dire que le taux correct du nombre d'expressions, d'idiomes et d'OOV dévoilé par notre système est plus élevé que pour les trois autres systèmes.

5.4 Résultat obtenus sur le grand corpus des médias sociaux

Nous utilisons un corpus qui contient au total 50 134 hanzi[49]. Au tableau 5.14, nous constatons que notre système a trouvé 20 987 mots, dont 3798 expressions, 145 idiomes, 786 pinyin, 2798 OOV et 489 symboles.

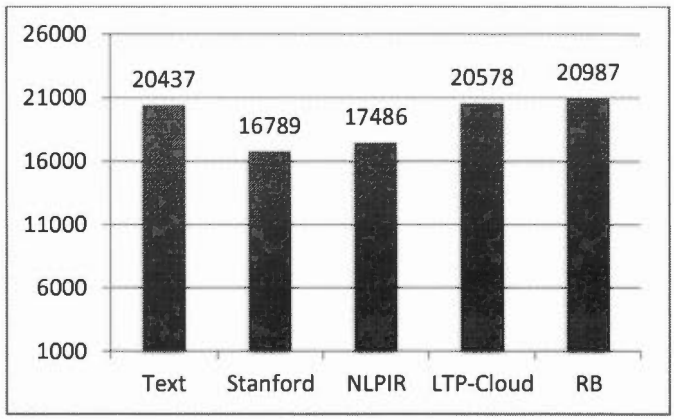
[49] Ce corpus vient du site www.nlpir.org, et la période du micro-blog SINA va du 12 octobre 2009 au 14 février 2010.

Corpus Médias Sociaux	Symbole	0	-	0	-	-	0
	OOV	2 378	-	0,7	-	-	2 789
	Pinyin	0	-	0	-	-	0
	Idiome	147	-	0,241	-	-	137
	Expression	2 478	-	0,168	-	-	2 648
	Mot	16 789	0,432	0,355	0,390	0,303	17 486
		Nombre	Précision	Rappel	F-mesure	Riv	Nombre
		Segmenteur Stanford					NLPIR (ICTCLAS)

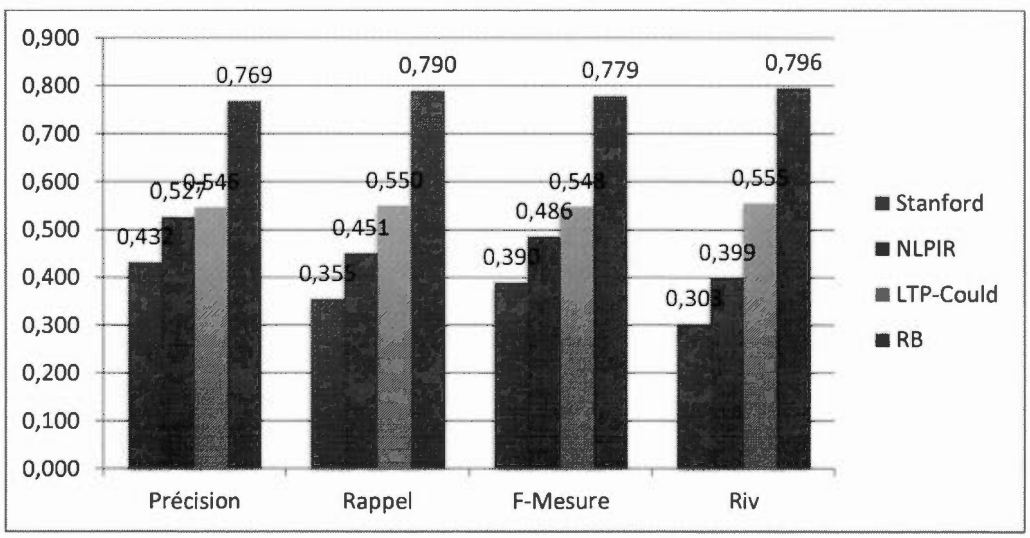
-	0	-	-	0	-	0	-	-
-	0,796	-	-	2 568	-	0,514	-	-
-	0	-	-	0	-	0	-	-
-	0,347	-	-	148	-	0,375	-	-
-	0,271	-	-	3 025	-	0,324	-	-
0,527	0,451	0,486	0,399	20 578	0,546	0,55	0,548	0,555
Précision	Rappel	F-mesure	Riv	Nombre	Précision	Rappel	F-mesure	Riv
				LTP-Cloud				

Notre système	Nombre	20 987	3 798	145	786	2 798	489
	Précision	0,769	-	-	-	-	-
	Rappel	0,790	0,547	0,624	0,484	0,747	0,471
	F-mesure	0,779	-	-	-	-	-
	Riv	0,796	-	-	-	-	-

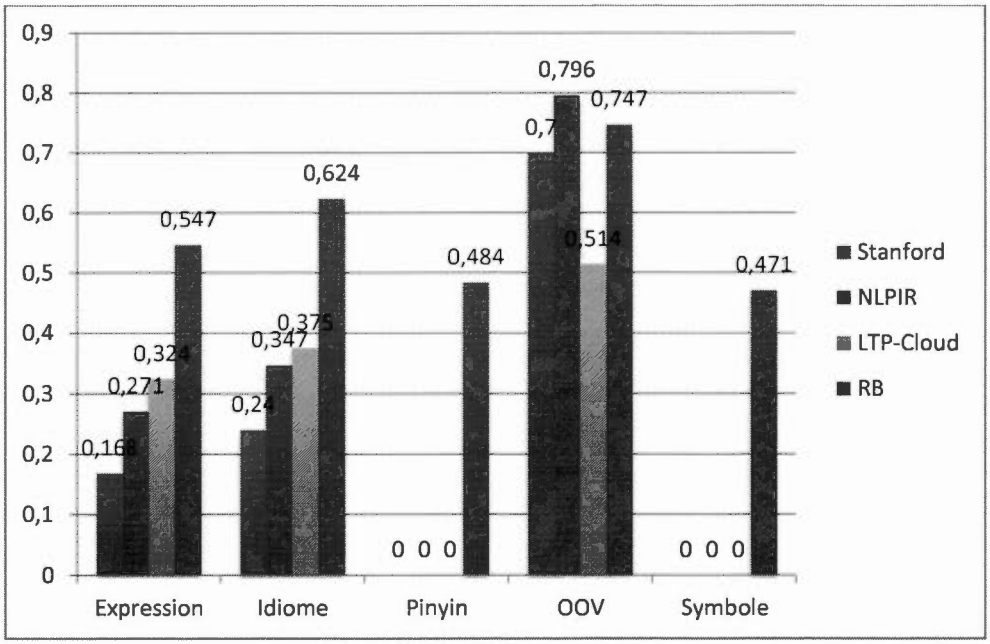
5.14 : Comparaison entre les systèmes pour le grand corpus des médias sociaux



5.15 : Nombre de mots segmentés par les trois outils et notre système sur le grand corpus des médias sociaux



5.16 : Statistique de la performance de la segmentation du grand corpus des médias sociaux



5.17 : Statistique des rappels de différents types de mots sur le grand corpus des médias sociaux

La figure 5.16 rapporte que notre système présente une haute valeur du F-mesure en combinant les deux segmentations.

CONCLUSION ET CONTRIBUTION

Conclusion

Depuis 1980, la première étape dans le traitement automatique de la langue chinoise est la segmentation du texte en mots. Bien que les systèmes de segmentation aient beaucoup évolué depuis cette époque, il n'existe toujours pas de système standard pour réaliser ce processus, le plus connu étant le segmenteur de Stanford. Le fait que différentes tâches demandent différents processus de segmentation signifie qu'il n'existe pas de segmentation universelle en chinois. D'ailleurs de nombreux travaux, comme le nôtre, sont basés sur le segmenteur de Stanford. Cependant, le segmenteur ce Stanford ne constitue qu'une des composants de notre système et peut donc être remplacé par d'autres segmenteurs. Les défaillances de la segmentation du texte chinois n'ont pas encore été complètement résolues, comme les noms propres des personnes et des lieux, les OOV et les ambiguïtés des mots.

Notre système est un système hybride. D'une part, il utilise le segmenteur de Stanford, une méthode basée sur un dictionnaire que nous avons enrichi d'un autre dictionnaire. Le système HowNet nous permet effectivement de regrouper ce second dictionnaire selon les relations sémantiques entre les mots, ce qui décupler la vitesse et la qualité de la segmentation. D'autre part, notre système utilise un réseau bayésien pour analyser les probabilités entre les caractères et les mots. De plus, notre système peut ajuster les probabilités entre les caractères et les mots en fonction des différentes influences pertinentes à la segmentation, comme par exemple la labellisation (POS) et la grammaire. Enfin, le résultat produit est mis à jour dans un nouveau dictionnaire et

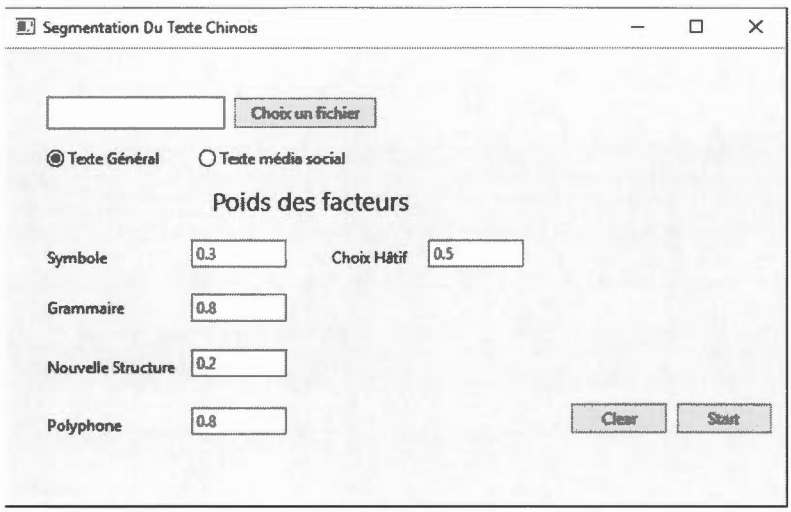
peut donc être réutilisé pour le prochain processus de segmentation, ce qui fait en sorte que les probabilités de nouveaux mots sont ajustées tout au long du traitement.

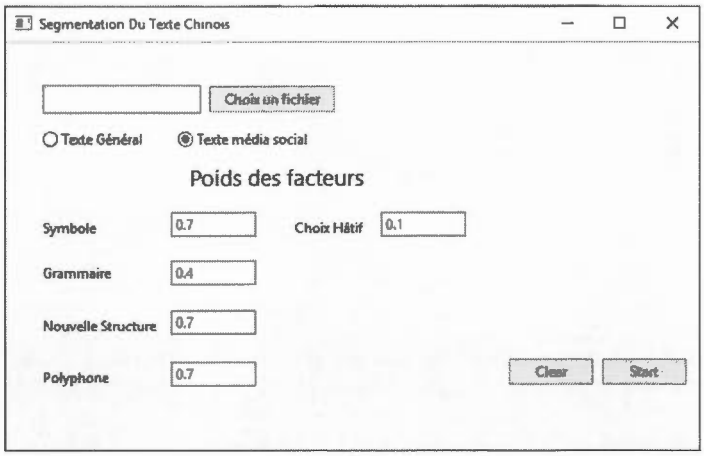
Par ailleurs, les travaux dans le domaine de la segmentation du pinyin rendent compte du fait que ce processus pose un problème identique à celui de la segmentation du hanzi. Les algorithmes de segmentation du pinyin exercent une grande influence sur la précision de la segmentation. Les plus importants sont l'adaptation maximale (MM), la méthode de la sélection des fréquences minimales (FWF) et la méthode *forward* de pinyin par pinyin.

Notre travail combine ces deux types de segmentations, celle du hanzi et celle du pinyin, ce qui permet d'intégrer le réseau bayésien au cœur de notre système. Nous utilisons la segmentation du pinyin pour améliorer la segmentation du hanzi. Le hanzi et le pinyin sont deux systèmes d'écriture différents de la langue chinoise, qui entrent toutefois en corrélation : le pinyin sert à prononcer le hanzi, et le hanzi est le symbole pour représenter le sens.

Une des particularités de notre système, qui en fait aussi son originalité, est la combinaison de deux différentes segmentations. Selon le flux de données auxquel on a fait référence précédemment, une phrase hanzi entre dans le système sous deux formes. En premier lieu, le système convertit la phrase hanzi vers le pinyin, puis la segmente par le segmenteur de Stanford, ce qui donne à la sortie des mots segmentés. Dans un second temps, la phrase en pinyin est segmentée par le segmenteur du pinyin, ce qui donne à la sortie des transcriptions de pinyin segmentées. Notre système considère trois critères importants pour une segmentation satisfaisante (la structure des phrases, les mots-clés et le *Semantic Chunk*) : il prendra en considération les probabilités de ces influences sur la segmentation du hanzi et sur celle du pinyin. Le réseau bayésien contribue à l'équilibre de ces deux segmentations successives, puisqu'il permet de trouver la bonne segmentation du hanzi correspondant avec exactitude à la bonne segmentation du pinyin.

Les influences de la segmentation du hanzi et du pinyin augmentent ou diminuent la probabilité de l'influence des critères. Selon les différents corpus, la valeur du poids des influences affecte aussi la probabilité de la segmentation du hanzi ou du pinyin ; c'est la raison pour laquelle nous pouvons ajuster le poids des influences par l'interface du logiciel. Par exemple, le corpus de journal présente une structure grammaticale assez standard, et donc, la valeur du poids d'influence grammatical est moins significative. Cependant, cette valeur est plus grande pour le corpus tiré des médias sociaux, parce que les commentaires qu'on y retrouve ne s'écrivent pas dans une grammaire standard. L'interface du système est présentée à la figure 5.18.





5.18: Interface du système

La figure 40 montre que nous pouvons ajuster chaque poids des influences et nous pouvons aussi choisir le type de corpus (le texte général ou le texte des médias sociaux). Par exemple, puisque nous allons traiter un corpus des médias sociaux et que ce type corpus contient peut-être beaucoup de symboles, nous pouvons augmenter le poids de l'influence des symboles. Pour les utilisateurs de notre système, nous pouvons fixer le poids de chaque influence à 0,5, valeur dont les utilisateurs peuvent les changer selon leurs besoins. Nous pourrions aussi développer un apprentissage automatique pour calculer automatiquement les meilleurs poids pour un type de corpus (basé par exemple sur de l'apprentissage machine ou un algorithme génétique). L'idée est alors de choisir d'abord un poids aléatoire, pour ensuite laisser le système modifier ceux-ci en fonction de la qualité de la segmentation produite, de façon à obtenir les poids les plus appropriés pour un type précis de corpus. Ce n'était l'objet de notre thèse que de développer une méthode pour déterminer les poids les plus appropriés pour chaque type de corpus, mais nous avons conçu notre système pour qu'une méthode d'apprentissage quelconque puisse aisément y être introduite à cette fin.

Dans notre système, nous considérons huit différentes influences pouvant agir sur la segmentation du hanzi et neuf influences qui peuvent agir sur la segmentation du pinyin. En ce qui concerne la segmentation du hanzi, nous prenons également en considération les relations sémantiques des hanzi en utilisant le HowNet et le *Semantic Chunk*. Quant à ce qui a trait à la segmentation du pinyin, nous considérons les relations sémantiques entre les caractères en utilisant toujours le *Semantic Chunk*. Chaque influence constitue un composant dans notre système ; aussi pouvons-nous la supprimer, la modifier ou la remplacer par d'autres influences (par exemple, pour d'autres types de corpus, d'autres types d'influences pourraient être découvertes).

Notre système peut être utilisé pour prétraiter un corpus en vue de sa traduction automatique statistique. L'utilisation de la segmentation du pinyin permet également d'utiliser notre système pour effectuer la vérification orthographique des textes chinois.

Limites et perspectives

Dans cette recherche, nous avons introduit un système bayésien de segmentation basé sur deux segmenteurs, celui de Stanford et celui du pinyin. Toutefois, en raison de la taille du corpus, le temps nécessaire pour calculer les 5-grams et la probabilité entre les mots est un long processus. Nous souhaitons développer notre système pour qu'il puisse travailler en parallèle, et donc calculer les probabilités entre les mots des différentes phrases en même temps.

Enfin, pour une évaluation de notre système sur un corpus de médias sociaux de 2 millions de caractères, nous souhaitons participer à un Bakeoff nous se basant sur les travaux de Zhao et al. (2010) et de Duan et al. (2012), car il serait bien intéressant de vérifier le résultat de notre système de segmentation sur 2 millions de caractères. Pour l'instant, toutefois, nous ne disposons pas d'une machine suffisamment puissante

pour le faire. Nous avons pu quand même montrer que notre système de segmentation est plus puissant que les autres trois systèmes de segmentation sur un corpus l'ordre de 20 000 mots.

Nous sommes convaincus que notre recherche pourrait être appliquée au développement de n'importe quelle application de traitement de la langue chinoise. D'une façon générale, le système pourrait être intégré à de nombreuses applications, tel que la traduction automatique, la segmentation des textes du domaine, etc. D'autres chercheurs pourraient également fusionner de nouveaux composants au système proposé, de manière à en optimiser la performance de la segmentation.

BIBLIOGRAPHIE

Ann, B., & Patrick, N. (1999). *Les réeasux bayésiens.* Édition Eyrolles.

Boutilier, C., Friedman, N., Goldszmidt, M., & Koller, D. (1996). Context-specific independence in Bayesian networks. *in processings of the 12th Conference on Uncertainty in Artifical Intelligence*, 115-123.

Cai, D., & Zhao, H. (2016). Neural Word Segmentation Learing for Chinese. *In Proceedings of ACL.* Berlin, Germany.

Chang, J., & Su, K. (1997). An unsupervised iterative method for Chinese new lexicon extraction. *International Journal of Computational Linguistics & Chinese Language Processing*.

Chen, K., & Liu, S. (1992). Word identification for Mandarin Chinese sentences. *Proceedings of the 14th conference on Computational linguistics, 1*.

Chen, M., Chang, B., & Pei, W. (2014). A joint Model for unsupervised Chinese Word Segmentation. *In Proceedings of the 2014 Conference on Empirical methodes in natural Language processing (EMNLP)*, (pp. 854-863). Doha, Qatar.

Chen, X., Qiu, X., & Huang, X. (2017). A Feature-Enriched Neural Model for Joint Chinese Word Segmentation and Part-of-Speech Tagging. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*, (pp. 3960-3966).

Chen, X., Qiu, X., Zhu, C., & Huang, X. (2015). Gated recursive neural network for chinese word segmentation. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, (pp. 1744-1753). Beijing, China.

Chen, Z., & Lee, K. (2000). A New Statistical Approach to Chinese Pinyin input. *Proceeding of the 38th Annua Meeting of the Association for Computational Linguistics*, (241-247).

Dai, L., Liu, B., Xia, Y., & Wu, S. (2008). Measuring Semantic Similarity between Words Using HowNet. *In proceedings of 2008 International Conference on*

Computer Science and Information Technology ICCSIT'08 (pp. 601-605). Washington, USA: IEEE Computer Scoiety.

Dang, Q., & Valette, M. (2017). Analyse sémantique du discours écologique relatifs au (wù maí),«brouillard de pollution» en Chine. *Actes des 9èmes Journées Internationales de la Linguistique de corpus*, (pp. 153-156). Grenoble, France.

David, D.-P., & Marti, A.-H. (1997). Adaptive multilingual sentence boundary disambiguation. *Computational Linguistics. 23 (2)*, (pp. 241-267). Cambridge, MA, USA: MIT Press.

d'Avignon, G., & Sauvageau, M. (1994). L'aide multicritère à la décision : un cas d'intégration de critères techniques, économiques et environnementaux à Hydro-Québec., (p. 17). Quebec.

Ding, Z.-G., Zhang, Z., & Li, J. (2009). Improvement on reverse directional maximum matching method based on hash structure for Chinese word segmentaion. *Computer Engineering and Design, 29*, (pp. 3208-3211, 3265).

Dong, Q., Hao, C., & Dong, Z. (2003). HowNet-Based Chinese Chunk Extractor (En Chinois). *In proceedings of the 7th Joint national Conference on Computational Linguistics*, (pp. 234-239). Haerbin, China.

Dong, Z., Dong, Q., & Hao, C. (2010). HowNet and Its Computation of Meaning. (pp. 53-56). Stroudsburg, PA, USA: Accosiation for Computational Linguistics.

Dong, Z.-D. (1999). Bigger Context and Better Understanding -- Expectation on Future MT Technology., (pp. 17-25).

Dong, Z.-D., & Dong, Q. (2006). *HowNet and the Computation of Meaning.* Singapore: World Scientific Press.

Duan, H., Sui, Z., Tian, Y., & Li, W. (2012). The CIPS-SIGHAN CLP 2012 Chinese Word Segmentation on MicroBlog Corpora Bakeoff. *In Proceedings of the Second CPS-SIGHAN Joint Conference on Chinese Language Processing*, (pp. 35-40). Tianjin, China.

Éric, P., & Jacques, B. (2007). *Le raisonnement bayésien : Modélisaton et inférene.* Springer.

Fenton, N., & Neil, M. (1999). Making Decisions : Bayesian Nets and MCDA. *Computer Science Departement, Queen Mary and Westfield college*.

Freitas, L., & Meng, C. (2013). *Pinyin to Hanzi Conversion with Hidden Markov Models.*

Fu, Z., & Delcroix, V. (2011). Bayesian network based on the method of ahp for making decision. *The IEEE Joint International Information Technology and Artificial Intelligence Conference ITAIC*, (pp. 223-227).

FU, Z., & Pierre, P. (2014, 13 Mai). Apprentissage non supervisé, basé sur réseau bayésien, de nouvelles unités lexicales pour améliorer la performance de la segmentation du texte chinois. *82e du Congrès de l'Acfas, Colloque 635 - Langues naturelles, informatique et sciences cognitives*.

Gan, K.-W., & Wong, P.-W. (2000). Annotating information structures in Chinese texts using HowNet. *In Proceedings of the second workshop on Chinese language processing: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics LCPW'00, 12*, (pp. 85-92).

Gao, J., Wang, H.-F., Li, M., & Lee, K.-F. (2000). A unified Approach to Statiscal Language Modeling for Chinese. *Proc. ICASSP, III*, (pp. 1703-1706).

He, S., He, N., Cen, S., & Lu, J. (2012). Semi-supervised Chinese Word Segmentation for CLP2012. *In Proceedings of the Second CPS-SIGHAN Joint Conference on Chinese Language Processing*, (pp. 79-84). Tianjin, China.

Huang, S.-L., Chung, Y.-S., & Chen, K.-J. (2008). E-HowNet: the expansion of HowNet.

Jensen, F. (2001). *Bayesian Networks and Decision Graphs.* Springer.

Jia, Z., Wang, P., & Zhao, H. (2013). Graph Model for Chinese Spell Checking. *Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing (SIGHAN-7)*, (pp. 88-92). Nagoya, Japan.

Jiang, W., Wang, T., Lin, T., Wang, F., Cheng, W., Liu, X., . . . Zhang, W. (2012). A rule based Chinese spelling and grammar detection system utility. *International Conference on System Science and Engineering (ICSSE)*, (pp. 437-440).

Jiao, L., & Peng, Y. (2012). Research and implementation of social hot topic detection system based on micro-blog. *In Proceedings of the 2012 IEEE 2nd International Conference on Cloud Computing and intelligent Systems (CCIS)*.

Langley, P., Iba, W., & Thompson, K. (1992). An analysis of Bayesian classier. *Proc. NCAI*, (pp. 223-228).

Leonardo, B. (2004). Chinese Text Word- Segmentation Considering Semantic Links among Sentences. *INTERSPEECH - ICSLP*.

Levy, R., & Manning, C. (2003). Is it harder to parse Chinese, or the Chinese Treebank? *In Proceedings of the 41st Annuel Meeting on Association for Computational Linguistics ACL'03, 1*, (pp. 439-446). Stroudsburg, PA, USA.

Li, D., & Peng, D. (2011). Spelling Correction for Chinese Language Based on Pinyin-Soundex Algorithm. *International Conference on Internet Technology and Applications (iTAP)*, (pp. 1-3).

Li, D., Ma, Y., & GUO, J. (2009). Words Semantic Orientation Classification Based On HowNet. *The Journal of China Universities of Posts and Telecommunications, 16*(1), (pp. 106-110).

Li, J., Zhou, G., & Ng, T.-H. (2010). Joint Syntactic and Semantic Parsing of Chinese. *In Proceedings of ACl 2010*, (pp. 1108-1117).

Li, X., Rayner, K., & Cave, R. K. (2009). On the segmentaion of chinese words during reading. *Cgnitive Psychology. 58*, (pp. 525-552). ELSEVIER.

Li, Z. (2011). Parsing the Internal Structure of Words: A New Paradigm for Chinese Word Segmentation. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies.*

Liang, N. (1987). Shumian hanyu zidong fenci xitong - CDWS (an automatic segmentation system for written Chinese - CDWS) (en Chinois). *Journal of Chinese Information Processing, 1(2)*, (pp. 44-52).

Liu, D., Fang, W., Zhou, H., & Li, Y. (2009 Aug). Bigram Chinese Word Segmentation by Viterbi Algorithm. *2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery.* (PP. *14-16)* . Tianjin, China: IEEEXpolre.

Liu, Q., & Li, S. (2002). Word Similarity Computing based on HowNet (En Chinois). *In proceedings of the 3rd Chinese Lexical Semantics Workshop.* Taibei, China.

Liu, X., Cheng, F., Luo, Y., Duh, K., & Matsumoto, Y. (2013). A Hybird Chinese Spelling Correction Using Language Model and Statistical Machine Translation with Reranking. *In Proceedings of the 7th SIGHAN Workshop on Chinese Language Processing, SIGHAN-7*, (pp. 54-58). Nagoya, Japan.

Liu, Y., & Li, W. (2014). Research of reverse Backtracking Matching Algorithm for Chinese Word Segmentation. *jornal of applied science and engineering innovation, 1*(3).

Liu, Y., Che, W., Guo, J., Qin, B., & Liu, T. (2016). Exploring Segment Representations for Neural Segmentation Models. *In Proceedings of the 25th International joint Conference on Artificial Intelligence IJCAI-16.*

Liu, Y., Tan, Q., & Shen, X.-K. (1994). The Word Segmentation Rules and Automatic Word Segmentation Methods for Chinese Information Processing (en Chinois). *Qing Hua University Press and guang Xi Science and Technology Press*, (p. 36).

Liu, Y., Zhang, M., Che, W., Liu, T., & Deng, Y. (2012). Micro blogs Oriented Word Segmentation System. *In Proceedings of the Second CPS-SIGHAN Joint Conference on Chinese Language Processing*, (pp. 85-99). Tianjin, China.

Lu, X. (2007). A Hybrid Model for Chinese Word Segmentation. *LDV-Forum 2007, Band 22 (1)*, (pp. 71-88).

Lucas, F., & Cynthia, M. (2013). Pinyin to Hanzi Conversion with hidden Markov Models. *Projet final.*

Ma, J., & Hinrichs, E. (2015). Accurate Linear-Time Chinese Word Segmentation via Embedding Matching. *In Proceedings of the 53rd Annual Meeting of the Association for Conputational Linguistics and the 7th International Joint Conference on Natural Language Processing*, (pp. 1733-1743).

Ma, Y. (1996). Jiyu pinjiade hanyu zidong fenci xitongde yanjiu yu shixian (Research and Implementation Based on Evaluation of Chinese Automatic Segmentation System) (en Chinois). *Language Information Processing*, (pp. 2-36).

Magistry, P. (2012). Segmentation Non Supervisée : Le Cas du Mandarin. *Actes de la conférence conjointe JEP-TALN-RECITAL 2012, 3*, (pp. 1-13). Grenoble, France.

Mark, J., & Katherine, D. (2010). Unsupervised Phonemic Chinese Word Segmentation using Adaptor Grammars. *In Proceedings of the 23th International Conference on Computional Linguistics, Coling 2010*, (pp. 528-536). Beijing, China.

Mo, J.-W., Zheng, Y., Shou, Z.-Y., & Zhang, S.-L. (2013). Improved Chinese word segementation method based on dictionary. *Computer Engineering and Design.*

Mochihashi, D., Yamada, T., & Ueda, N. (2009). Bayesian Unsupervised Word Segmentation with Nested Pitman-Yor Language Modeling. *In ACl-IJCNLP'09: Proceedings of the Joint Conference of the 47th Annual meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 1*, (pp. 100-108). Morristown, NJ, USA.

Nie, J.-Y., Hannan, M.-L., & Jin, W. (1995). Combining dictionary, rules and statistics in unknown word detection and segmentation of Chinese. *Computer Processing of Oriental Languages, 9 (2)*, (pp. 125-144).

Peng, F., & Schuurmans, D. (2001). Self-supervised Chinese word segmentation. *Advances in Intelligent Data Analysis.*

Qian, X., & Liu, Y. (2012). Joint Chinese Word Segmentation, POS Tagging, and Parsing. *In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, (pp. 501-511).

Qiu, X., Yin, L., Wu, S., & Huang, X. (2015). Overview of the NLPCC 2015 Shared task: Chinese Word Segmentation and POS Tagging for Micro-blog Texts.

Proceedings 4th CCF Conference, Natural Language Processing and Chinese Computing, NLPCC, (pp. 541-549). Nanchang, China.

Sedki, K., Delcroix, V., Lepoutre, F.-X., Adam, E., Maquinghen-Godillon, A.-P., & Ville, I. (2010). Bayesian Network Model for Decision problems. *the 18th International Conference INTELLIGENT INFORMATION SYSTEMS 2010, IIS 2010.*

Sha, Y., Xia, M., Jiang, H., & Wang, X. (2012). Word Semantic Orientation Algorithm Based on Dynamic Standard Word Set for Multi-dimain. *Journal of Computational Information Systems. 8:3*, (pp. 1001-1009). http://www.Jofcis.com.

Sharon, G., Thomas, L.-G., & Mark, J. (2006). Contextual Dependencies in Unsupervised Word Segmentation. *In Proceeding of the 44th Annual Meeting of the Association for Computational Linguistics.*

Song, P., & al. (2012). A Pointillism Approach for Natural Language Processing of Social Media. *In IEEE International Conference on Natural Language Processing and Knowledge Engineering.*

Spirtes, P., & al. (1993). *Causation Prediction and Search.* New York: Springer-Verlag.

Sproat, R., & Shih, C. (1990). A statistical method for finding word boundaries in Chinese text. *Computer Processing of Chinese and Oriental Languages*, 4:336-351.

Stich, T. (2004). *Bayesian networks and structure learning*. Récupéré sur Diploma Thesis, Computer Science and Engineering, University of Mannheim: http://66.102.1.104/scholar?hl=en&lr=&q=cache:j36KPn-8hWroJ:www.timostich.de/resources/thesis.pdf

Sun, M., Shen, D., & Tsou, B. (1998). Chinese Word Segmentation without Using Lexicon and Hand-crafted Training Data . *Proceedings of the 17th international conference on Computational linguistics, 2.*

Sun, M.-S., & Benjamin, K.-T. (1995). Ambiguity resolution in Chinese word segmentation. *In Proceedings of the 10th Asia Conference on Language, Inforamtion and Computation*, (pp.121-126).

Sun, W., & Xu, J. (2011). Enhancing Chinese Word Segmentation Using Unlabeled Data. *In Proceedings of the 2011 Comference on Emprical Methods in Natral language Processing* (pp. 970-979). Edinburgh, Scotland, UK: Association for Computational Linguistics.

Tanabe, T., Takahashi, M., & Shudo, K. (2014, 11). A lexicon of multiword expressions for linguistically precise wide-coverage natural language processing. *Computer Speech & Language, 28*(6), (pp. 1317-1339).

Tang, J., Wu, Q., & Li, Y. (2015). An Optimization Algorithm of Chinese Word Segmentation Based on Dictionary. *In Proceedings of the 2015 International Conference on Network and Information Systems for Computers (ICNISC)*, (pp. 259-262).

Underhill, P.-A. (2013). *A Companion to Chinese archaeology.* Chichester, West Sussex; Malden (Mass.), Wiley-Blackwell.

Wang, C.-Y. (2002). Knowledge-based Sense Pruning using the HowNet: an Alternative to Word Sense Disambiguation. *A thesis Submitted to The Hong Kong University of Science & Technology.*

Wang, F.-L., Deng, X., & Zou, F. (2006). Towards Unified Chinese Segmentation Algorithm. *Proceedings 5th Edition of the International Conference on Language Ressources and Evaluation*, (pp. 379-384). Genoa, Italy.

Wang, H., Zhu, J., Tang, S., & Fan, X. (2011). A New unsupervised Approach to Word Segmentation. *Computional Linguistics, 37*(3), (pp. 421-454).

Wang, L., & Yu, S. (2010). Construction of a Chinese Idiom Knowledge Base and Its Applications. *Proceedings of the Multiword Expressions: From Theory to Applications MWE*, (pp. 11-18).

Wang, L., Yu, S., Zhu, X., & Li, Y. (2012). Chinese idiom knowledge base for chinese information processing. *Proceeding CLSW'12 Proceedings of the 13th Chinese conference on Chinese Lexical Semantics*, (pp. 302-310).

Wang, L., Yu, S., Zhu, X., Lo, F., Sunaoka, K., & Kang, B. (2013). Principle and New Development of Constructing Chinese Idiom Knowledge Base (En Chinois). *In Proceedings of 4th Internation Symposium on Chinese Classics Digitization.* Beijing, China.

Wang, P., Qian, Y., Zhao, H., Soong, F.-K., He, L., & Wu, K. (2016). Learing Distributed Word Representations for Bidirectional lstm Recurrent Neural Network. *In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.*

Wang, X.-L., Wang, K.-Z., Li, Z.-R., & Bai, X.-H. (1989). Fewest Word Matching problem and its solution (en Chinois). *Chinese Science Bulletin, 34*(13), (pp. 1031-1032).

Wei, R. (2009). The state of new media technology research in china: a review and critique. *Asian Journal of Communication, 19*(1), (pp. 116-127).

Wen, S., Li, Z., & Li, J. (2014). Enhance Social Context understanding with Semantic Chunks. *Proceedings 3rd CCF Conference, Natural Language Processing and Chinese Computing, NLPCC*, (pp. 251-262). Shenzhen, China.

Wong, P.-K., & Chan, C. (1996). Chinese Word Segmentation Based on Maximum Matching and Word Binding Force. *In Proceedings of the 16th Conference on Computational Linguistics (COLING'96), 1*, (pp. 200-203).

Wu, L.-C. (2010). Outils de Segmentation du Chinois et textométrie. *Actes de la 12e Rencontres des Étudiants Chercheurs en informatique pour le Traitement Automatiques des Langues (RECITAL'2010).* Montréal, Canada.

WU, Z. (2011). A Cognitive Model of Chinese Word Segmentation for Machine Translation. *Translators' Jornal, 56*(3), (pp. 631-644).

Xia, F. (2000). The Part-Of-Speech Tagging Guidelines for the Penn Chinese Treebank (3.0).

Xu, J., Gao, J., Toutanova, K., & Ney, H. (2008). Bayesian Semi-Supervised Chinese Word Segmentation for Statical Machine Translation. *In Proceedings of the 22nd International Conference on Computational Linguistics*, (pp. 1017-1024). Manchester, United Kingdom.

Xue, N. (2003). Chinese Word Segmentation as Character Tagging. *Computational Linguistics and Chinese Language Processing.*

Xue, N., & Converse, S. (2002). Combining classifiers for Chinese word segmentation. *Proceedings of the 1st {SIGHAN} Workshop on Chinese Language Processing.*

Xue, N., & Shen, L. (2003). Chinese Word Segmentation as-LMR Tagging. *Proceedings of the second (SIGHAN) workshop on Chinese language processing, 17.*

Yang, H., & Zong, C. (2014). A Global Generative Model for Chinese Semantic Role Labeling. *Proceedings 3rd CCF Conference, Natural Language Processing and Chinese Computing, NLPCC*, (pp. 1-12). Shenzhen, China.

Yang, S., & al. (2012). Spell Checking for Chinese. *In International Conference on Language Resources and Evaluation*, (pp. 730–736).

Yang, S., Zhao, H., Wang, X., & Lu, B. (2012). Spell Checking for Chinese. *International Conference on Language Resources and Evaluation*, (pp. 730-736). Istanbul, Turkey.

Yuan, Z., & Purver, M. (2012). Prediction Emontion Labels for Chinese Microblog Texts. *In Procedding of the 1st International Workshop on Sentiment Discovery from Affective Data (SDAD)*, (pp. 40-47). Bristol, UK.

Zhang, G., & Simon, H.-A. (1985). STm Capacity for Chinese Words and idioms: Chunking and the Acoustical Loop Hypothesis. *Memory & Cognition, 13*, (pp. 193-201).

Zhang, H.-P., YU, H.-K., Xiong, D.-Y., & Liu, Q. (2003). HHMM-Based Chinese Lexical Analyzer ICTCLAS. *In Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing*, (pp. 184-187).

Zhang, J., Huang, D., Han, X., & Wang, W. (2012). Rules-based Chinese Word Segmentation on MicroBlog for CIPS-SIGHAN on CLP2012. *In Proceedings of the Second CPS-SIGHAN Joint Conference on Chinese Language Processing*, (pp. 74-78). Tianjin, China.

Zhang, K., Sun, M., & Zhou, C. (2012). Word Segmentation on Chinese Micro-Blog Data with Linear-time Incremental Model. *In Second CIPS-SIGHAN Joint Conference on Chinese Language Processing*, (pp. 41-46).

Zhang, M., Zhang, Y., & Fu, G. (2016). Transition-based Neural Word Segmentation. *In Proceedings of the 54nd Annual Meeting of the Associationg for computational Linguistics.*

Zhang, R., Yang, G., & Yan, X. (2012). Hownet-based Semantic Analysis Model for General Unknown Chinese Words. *Coumputet Applications and Software, 29 (8)*, (pp. 126-130).

Zhang, X., & al. (2014). A Distributed Approach For Chinese Micro-blog Hot Topic Detection. *International Conference on Logistics, Engineering, Management and Computer Science (LEMCS 2014).*

Zhao, H., & Kit, C. (2008). An Empirical Comparison of Goodness Measures for Unsupervised Chinese Word Segmentation with a Unified Framework. *In proceedings of IJCNLP 2008.*

Zhao, H., & Liu, Q. (2010). The CIPS-SIGHAN CLP 2010 Chinese Word Segmentation Bakeoff. *In Proceedings of the First CPS-SIGHAN Joint Conference on Chinese Language Processing*, (pp. 199-209). Beijing, China.

Zhao, H., Huang, C.-N., & Li, M. (2006). An Improved Chinese Word Segmentation System with Conditional Random Field. *In Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, 1082117.*

Zheng, C., Liu, W., & Feng, Z. (2002). A New Statistical Approach to Personal Name Extraction. *Proceedings of the 19th International Conference on Machine Learning*, (pp. 67-74).

Zheng, X., Chen, H., & Xu, T. (2013). Deep Learning for Chinese Word Segmentation and POS Tagging. *In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, (pp. 647-657).

Zheng, Y., Li, C., & Sun, M. (2011). CHIME: An Efficient Error-Tolerant Chinese Pinyin Input Method. *In Proceedings of the Twenty-Sceond International Joint Conference on Artificial Intelligence IJCAI'11. 3*, (pp. 2551-2556). AAAI Press.

Zheng, Y., Xie, L., Liu, Z., Sun, M., Zhang, Y., & Ru, L. (2011). Why Press Backspace? Understanding User Input Behaviors in Chinese Pinyin Input Method. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Techologies* (pp. 485-490). Portland, Oregon, USA: Association for Computational Linguistics.

Zhong, K., Zhou, X., Li, H., & Yuan, C. (2012). Cascaded Chinese Weibo Segmentation Based on CRFs. *Proceeding of the Second CIPS-SIGHAN Joint Conference on Chinese language Processing*, (pp. 69-73). Tianjin, China.

Zhou, Q., Drabek, F.-E., & Ren, F. (2002). Annotation the Functional Chunks in Chinese Sentences. *In Proceeding of the 3rd International Conference on Language Resources and Evaluation* (pp. 731-738). Paris, France: European Language Resources Association.

Zhu, Z., & Sun, J. (2013, 08 01). Improved Vocabulary Semantic Similarity Calculation Based On HowNet (En Chinois)., *33*, (pp. 2276-2279,2288).