Segmentation non supervisée : le cas du mandarin

Pierre Magistry
Alpage, INRIA Paris–Rocquencourt & Université Paris Diderot

RÉSUMÉ

Dans cet article, nous présentons un système de segmentation non supervisée que nous évaluons sur des données en mandarin. Notre travail s'inspire de l'hypothèse de Harris (1955) et suit Kempe (1999) et Tanaka-Ishii (2005) en se basant sur la reformulation de l'hypothèse en termes de variation de l'entropie de branchement. Celle-ci se révèle être un bon indicateur des frontières des unités linguistiques. Nous améliorons le système de (Jin et Tanaka-Ishii, 2006) en ajoutant une étape de normalisation qui nous permet de reformuler la façon dont sont prises les décisions de segmentation en ayant recours à la programmation dynamique. Ceci nous permet de supprimer la plupart des seuils de leur modèle tout en obtenant de meilleurs résultats, qui se placent au niveau de l'état de l'art (Wang *et al.*, 2011) avec un système plus simple que ces derniers. Nous présentons une évaluation des résultats sur plusieurs corpus diffusés pour le *Chinese Word Segmentation bake-off II* (Emerson, 2005) et détaillons la borne supérieure que l'on peut espérer atteindre avec une méthode non-supervisée. Pour cela nous utilisons ZPAR en apprentissage croisé (Zhang et Clark, 2010) comme suggéré dans (Huang et Zhao, 2007; Zhao et Kit, 2008)

ABSTRACT

Unsupervized Word Segmentation

In this paper, we present an unsupervised segmentation system tested on Mandarine Chinese. Following Harris's Hypothesis in Kempe (1999) and Tanaka-Ishii (2005) reformulation, we base our work on the Variation of Branching Entropy. We improve on (Jin et Tanaka-Ishii, 2006) by adding normalization and Viterbi-decoding. This enables us to remove most of the thresholds and parameters from their model and to reach near state-of-the-art results (Wang *et al.*, 2011) with a simpler system. We provide evaluation on different corpora available from the Segmentation bake-off II (Emerson, 2005) and define a more precise topline for the task using cross-trained supervised system available off-the-shelf (Zhang et Clark, 2010; Zhao et Kit, 2008; Huang et Zhao, 2007)

MOTS-CLÉS : Apprentissage non-supervisé, segmentation, chinois, mandarin.

KEYWORDS: Unsupervized machine learning, segmentation, Mandarin Chinese.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 3: RECITAL, pages 1–13,
Grenoble, 4 au 8 juin 2012. ©2012 ATALA & AFCP

1 Introduction

Pour la plupart des langues utilisant l'alphabet latin, un découpage sur les espaces est une bonne approximation d'une segmentation en unités lexicales. L'écriture chinoise en revanche ne délimite pas ces unités par la typographie. Seules les marques de ponctuation indiquent une partie des frontières entre les unités lexicales qui peuvent être formées d'un ou plusieurs caractères chinois. L'étape de *tokenisation*, préalable à beaucoup de systèmes d'analyse automatique est de ce fait plus délicate. Pour les langues sans caractère d'espacement ou équivalent, on parle d'étape de segmentation en mot.

De nombreux systèmes de segmentation par apprentissage supervisé ont été proposés mais ils requièrent des corpus segmentés manuellement. Ceux-ci sont souvent spécifiques à un genre, un domaine ou une variété de mandarin et en l'absence d'un consensus sur la définition de ce qu'est un « mot », ils suivent des guides d'annotations qui divergent.

Les systèmes supervisés atteignent aujourd'hui des résultats satisfaisants lorsque le corpus approprié pour l'entraînement est disponible. Cependant, si l'on veut faire face à une plus grande diversité en genres et en domaines ou répondre à des questions plus théorique sur la caractérisation formelle des unités de langue, s'intéresser aux approches non-supervisées nous semble nécessaire.

De plus, il est important de souligner que le système que nous présentons ici n'est pas spécifique au mandarin, ni à la segmentation de séquence de caractères chinois en « mots ». Des expérimentations sur d'autres langues et à partir d'autres unités initiales (lettre, phonème, mot orthographique) sont en cours et donnent des résultats prometteurs. En l'état actuel de l'avancement de nos travaux, nous ne pouvons fournir d'évaluation complète que pour le mandarin. Nous limitons donc notre présentation à cette langue.

Cette article est organisé ainsi : après une courte présentation de l'état de l'art à la section 2, nous détaillons les méthodes et les difficultés d'évaluation des systèmes de segmentation non-supervisés à la section 3. Les sections 4 à 5 présentent le fonctionnement notre système et la section 6 les résultats obtenus.

2 État de l'art

Les systèmes de segmentation non supervisés reposent généralement sur un des trois types de mesures suivantes ou une combinaison des trois : le niveau de cohésion des unités obtenues (par exemple en utilisant l'information mutuelle, comme dans (Sproat et Shih, 1990)) ; le degré de séparation des unités obtenues (par exemple la diversité des contextes, (Feng *et al.*, 2004)) ou la probabilité d'une segmentation étant donnée une chaîne (Goldwater *et al.*, 2006; Mochihashi *et al.*, 2009).

Dans un article publié récemment, Wang *et al.* (2011) présentent une méthode baptisee « *Evaluation, Selection, Adjustment*. » (*ESA*). Cette méthode combine cohésion et séparation en une mesure à maximiser sur l'ensemble d'une séquence. Ils utilisent ensuite les résultats de leur système pour en modifier les paramètres (essentiellement les comptes de n-grammes) et réitérer le processus 10 à 30 fois. Ils obtiennent ainsi les meilleurs résultats actuels en segmentation non-supervisée du mandarin.

Les principaux inconvénients de l'approche ESA sont d'une part le fait qu'il faille itérer le processus sur le corpus environ 10 fois avant d'atteindre des niveaux de performance satisfaisants, et d'autre part la nécessité d'avoir à fixer le paramètre de couplage entre mesure de cohésion et mesure de séparation. Empiriquement, on constate une corrélation entre ce paramètre et la taille du corpus, mais cette corrélation dépend de la façon dont sont traités les caractères latins et les chiffres arabes au cours des prétraitements. De plus, calculer cette corrélation et choisir la valeur optimale du paramètre en question (ce que les auteurs appellent le *proper exponent*) nécessite un corpus segmenté à la main, ce qui contredit le caractère non-supervisé de l'approche. Toutefois, si parmi les différents types de prétraitements pour lesquels ESA a été évalué on se réfère aux configurations qui se rapprochent des nôtres, les résultats de Wang *et al.* (2011) avec leur approche ESA se situent tous aux alentours de 0,80 de f-mesure sur les mots.

L'approche plus ancienne de Jin et Tanaka-Ishii (2006) ne repose que sur une mesure de séparation, elle-même directement insipirée par une hypothèse linguistique formulée par Harris (1955). Reformulée au moyen de la notion d'entropie de branchement (*Branching Entropy*, *BE*) par Tanaka-Ishii (2005) en suivant les travaux de Kempe (1999), cette hypothèse peut s'énoncer comme suit : si les séquences de graphèmes, phonèmes, ou autres produites par l'homme étaient aléatoires, on s'attendrait à ce que l'entropie de branchement d'une séquence (estimée à partir de *n*-grammes en corpus) décroisse lorsque la longueur de la séquence croît. Ainsi, la variation de l'entropie de branchement (*Variation of the Branching Entropy*, *VBE*) devrait être systématiquement négative. Lorsque l'on observe au contraire une VBE positive, l'hypothèse de Harris conduit à conclure que l'on se situe à une frontière d'unités linguistiques. C'est sur la base de cette hypothèse que Jin et Tanaka-Ishii (2006) proposent un système qui segmente dès que la BE croît (c'est-à-dire que la VBE est positive) ou lorsqu'elle atteint un certain maximum. Les auteurs fixent la longueur maximale des séquences calculées à 6 et lisent le corpus de gauche à droite et de droite à gauche. À chaque intervalle entre deux caractères, ils peuvent donc observer jusqu'à 12 valeurs desquelles ils conservent le maximum.

Le principal inconvénient de l'approche de Jin et Tanaka-Ishii (2006) est que les décisions de segmentation sont prises très localement [1] et ne dépendent pas des segmentations voisines. De plus, ce système repose lui aussi sur des paramètres, et notamment le seuil sur la VBE au dessus duquel le système décide de segmenter (dans leur système, il y a segmentation dès lors que VBE≥ 0). En théorie, on pourrait décider de segmenter dès lors que la BE ne décroît pas suffisamment, ou à l'inverse ne segmenter que si la VBE est non seulement positive mais même au dessus d'un certain seuil non nul. À cet égard, placer le seuil à la valeur 0 peut être considéré comme une valeur par défaut, mais reste un paramètre adaptable. Enfin, Jin et Tanaka-Ishii ne prennent pas en compte le fait que la VBE pour un n-gramme n'est pas forcément comparable *a priori* avec la VBE pour un m-gramme dès lors que $n \neq m$: une normalisation est ici nécessaire, comme le suggèrent notamment Cohen *et al.* (2002).

Faute de place, nous ne décrirons pas ici d'autres systèmes que ceux de Wang *et al.* (2011) et de Jin et Tanaka-Ishii (2006). Un état de l'art plus exhaustif peut être trouvé dans les articles de (Zhao et Kit, 2008) et de (Wang *et al.*, 2011).

Dans cet article, nous montrons que l'on peut corriger les inconvénients du modèle de Jin et Tanaka-Ishii (2006) et atteindre des niveaux de performance comparables à ceux de l'état de l'art, c'est-à-dire de Wang *et al.* (2011), le tout avec un système plus simple.

1. Dans sa thèse, Jin utilise l'auto-apprentissage et le paradigme de la *minimum description length (MDL)* pour pallier à ce problème.

	mots		caractères	
Corpus	en tout	différents	en tout	différents
Academia Sinica (AS)	5 449 698	141 340	8 368 050	6 117
City University of Hong Kong (CITYU)	1 455 629	69 085	2 403 355	4 923
Peking University (PKU)	1 109 947	55 303	1 826 448	4 698
Microsoft Research (MSR)	2 368 391	88 119	4 050 469	5 167

TABLE 1 – Taille des corpus utilisés

3 Problèmes d'évaluation

Dans cet article, afin de pouvoir nous comparer au système de Wang *et al.* (2011), nous nous évaluons sur les corpus du second Bakeoff international de segmentation du chinois (*Second International Chinese Word Segmentation Bakeoff*, Emerson, 2005). Ces corpus couvrent 4 guides de segmentation différents, développés au sein de 4 institutions distinctes : l'Academia Sinica (AS), la *City University* de Hong-Kong (CITYU), l'université de Pékin (PKU) et Microsoft Research (MSR).

Des informations sur la taille des corpus sont données au tableau 1. Dans le cadre du *Bakeoff*, Les détails du contenu n'étaient pas connus des participants. Mais on peut noter que le corpus de PKU est constitué d'extraits du Quotidien du Peuple, journal de Pékin. Le corpus de CITYU est extrait du LIVAC (T'sou *et al.*, 1997), aussi constitué de textes de presse, mais d'origines plus variées. Le projet du LIVAC cherchant à rendre compte des variantes géographique du mandarin, il inclut des articles provenant de Pékin, Hong-Kong, Singapour ou Taïwan. Le corpus de l'AS est un corpus équilibré, qui rend essentiellement compte de la variante du mandarin utilisée à Taïwan. Aucune description du contenu du corpus de MSR n'est disponible à notre connaissance.

L'évaluation de systèmes non-supervisés est une problématique en soi. Un consensus sur une définition précise de la notion de *mot* restant difficile à atteindre, différents guides d'annotation pour la segmentation en mots ont été proposés et appliqués à divers corpus. L'évaluation de systèmes de segmentation supervisés peut être réalisée sur n'importe quel corpus, indépendamment du guide d'annotation sous-jacent, pour peu que les données d'entraînement et les données d'évaluation soient cohérentes. Cependant, pour les systèmes non-supervisés, il n'y a aucune raison d'obtenir des résultats plus proches de l'un des guides existants que d'un autre, plutôt que des résultats se situant quelque part entre les différents guides. Huang et Zhao (2007) propose d'utiliser l'entraînement et l'évaluation croisés de systèmes de segmentation supervisés pour avoir un ordre d'idée sur le taux de désaccord entre guides d'annotation. L'idée est donc d'entraîner puis d'évaluer un système supervisé sur deux corpus respectant deux guides d'annotation distincts, et d'en tirer une approximation de leur désaccord. C'est également un moyen d'estimer une borne supérieur de ce que l'on est en droit d'attendre de la part d'un système non-supervisé, qui n'a pas de raison d'être plus proche d'un guide d'annotation que ne le sont les autres guides existants (Zhao et Kit, 2008). Nous avons reproduit ce type de mesures sur nos 4 corpus au moyen du système supervisé ZPAR (Zhang et Clark, 2010), et nous avons trouvé une cohérence moyenne similaire à celle obtenue par Huang et Zhao (2007), de l'ordre de seulement 0,84 (f-mesure), qui sera donc notre *topline*. Par ailleurs, il est généralement admis que segmenter chaque caractère individuellement est une *baseline* raisonnable, puisque près de la moitié des mots-formes dans un corpus segmenté à la main sont des unigrammes. Une telle baseline obtient

un f-score d'environ 0,35.

Ces évaluations globales peuvent être raffinées en décomposant les résultats en fonction de la longueur des mots. Les mots de longueurs différentes ont en effet des distributions très dissemblables. Les évaluations par longueur donnent les résultats suivants : sur les unigrammes, les f-scores se situent entre 0,81 et 0,90, similaires aux résultats globaux. Les résultats pour les bigrammes sont légèrement meilleurs (0,85–0,92), mais bien plus bas sur les trigrammes, descendant entre 0,59 et 0,79. Or, dans un texte en mandarin, la majorité des occurrences sont des mots unigrammes ou bigrammes, mais le lexique est principalement composé de bigrammes et de trigrammes. Ceci vient du fait que les unigrammes sont souvent des mots grammaticaux à haute fréquence, alors que les trigrammes sont souvent le résultat d'affixations plus ou moins productives. Pour cette raison, les résultats uniquement calculés sur les occurrences ne pâtissent pas énormément de mauvaises performances sur les trigrammes, même si une proportion significative du lexique est ainsi mal traitée.

Une autre difficulté concernant l'évaluation et la comparaison entre systèmes non-supervisés est de prendre en compte de façon équitable les prétraitements et les connaissances *a priori* qui sont fournies aux systèmes. Par exemple, Wang *et al.* (2011) utilise différents niveaux de prétraitement (qu'ils appellent *settings* et que nous appelerons « configurations »). Dans les configurations 1 et 2, Wang *et al.* (2011) essayent de ne pas se reposer sur la ponctuation et l'encodage des caractères (notamment la distinction entre caractères chinois et latins). Cependant, ils optimisent indépendamment leur paramètre pour chaque configuration. Nous considérons donc que leur système prend en compte le niveau de prétraitement qui est effectué sur les caractères latins et les chiffres romains, et sait donc à quoi s'attendre en la matière. Dans leur configuration 3, les auteurs ajoutent la connaissance de la ponctuation en tant que frontières de mots, et leur configuration 4 ajoute à cela un prétraitement des caractères latins et des chiffres arabes, ce qui conduit à des résultats plus significatifs, moins questionnables et plus convaincants.

Nous sommes plus intéressés par une réduction du travail humain que par le déploiement à tout prix d'un système strictement non-supervisé. Nous ne pensons donc pas utile de nous empêcher de procéder à quelques prétraitements simples, tels que ceux discutés ci-dessus : détection des ponctuations, des caractères latins et des chiffres arabes[2]. C'est la raison pour laquelle nos expériences correspondent aux configurations 3 et 4 de Wang *et al.* (2011), et c'est à elles que nous nous comparons, en appliquant notre système aux mêmes corpus.

4 Variation de l'entropie de branchement

4.1 Formulation

Notre système repose sur l'hypothèse de Harris (1955) et sa reformulation par Kempe (1999) et Tanaka-Ishii (2005). Définissons à présent les notions sous-jacentes à notre système.

Soit un n-gramme $x_{0..n} = x_{0..1}\, x_{1..2} \ldots x_{n-1..n}$ dont le contexte droit $\chi_{\rightarrow}$ contient tous les caractères observés à sa droite dans le corpus. Nous définissons son *entropie de branchement droite* (*Right*

2. De simples expressions régulières peuvent aussi être envisagées pour traiter les cas non-ambigus de nombres et de dates utilisant les sinogrammes

Branching Entropy, *RBE*) comme suit :

$$\begin{aligned} h_{\rightarrow}(x_{0..n}) &= H(\chi_{\rightarrow} \mid x_{0..n}) \\ &= -\sum_{x \in \chi_{\rightarrow}} P(x \mid x_{0..n}) \log P(x \mid x_{0..n}). \end{aligned}$$

L'*entropie de branchement gauche* (*Left Branching Entropy*, *LBE*) est définie de façon symétrique : si l'on note $\chi_{\leftarrow}$ le contexte gauche de $x_{0..n}$, sa LBE est définie par :

$$h_{\leftarrow}(x_{0..n}) = H(\chi_{\leftarrow} \mid x_{0..n}).$$

La RBE $h_{\rightarrow}(x_{0..n})$ peut être considérée comme l'entropie de branchement (BE) de $x_{0..n}$ au cours d'un parcours de gauche à droite, alors que la LBE est la BE de $x_{0..n}$ au cours d'un parcours de droite à gauche.

À partir, d'une part, de $h_{\rightarrow}(x_{0..n})$ et $h_{\rightarrow}(x_{0..n-1})$, et d'autre part de $h_{\leftarrow}(x_{0..n})$ et $h_{\leftarrow}(x_{1..n})$, nous définissons la *variation de l'entropie de branchement* (*Variation of Branching Entropy*, *VBE*) dans les deux directions comme suit :

$$\begin{aligned} \delta h_{\rightarrow}(x_{0..n}) &= h_{\rightarrow}(x_{0..n}) - h_{\rightarrow}(x_{0..n-1}) \\ \delta h_{\leftarrow}(x_{0..n}) &= h_{\leftarrow}(x_{0..n}) - h_{\leftarrow}(x_{1..n}). \end{aligned}$$

4.2 Observations intermédiaires

Après avoir reproduit les expériences de Jin et Tanaka-Ishii (2006), nous avons effectué une série d'observations des valeurs prises par l'entropie de branchement et ses variations. Certaines de ces observations ont motivé les modifications apportées au modèle que nous présenterons à la section suivante. Dans cette section, nous présentons les plus pertinentes de ces observations. Pour des questions de place disponible et pour éviter la redondance, les graphiques de cette section sont produits à partir du corpus de PKU uniquement.

4.2.1 Confirmation de l'hypothèse de Harris

Dans un premier temps, nous allons confirmer l'hypothèse de Harris sur nos données. Pour cela nous nous limitons à l'observation de la frontière droite des bigrammes de notre corpus (le choix des bigrammes étant motivé par leur représentativité tant en nombre d'occurences en corpus qu'en nombre d'entrées dans le lexique). Cette valeur est donc calculée pour chaque bigramme observé au moins deux fois dans le corpus. On affiche ensuite l'ensemble de ces valeurs sous forme d'une courbe de densité qui donne ainsi la répartition des valeurs prises par la variation d'entropie (c'est à dire les '$\delta h_{\rightarrow}(x_{0..2})$). On distingue ensuite de l'ensemble de tous les bigrammes ceux qui sont considérés comme des mots par l'annotation manuelle de ceux qui ne le sont pas. Le résultat est présenté figure 4.1.1. On observe que les mots valides (qui forment une très petite proportion de l'ensemble des bigrammes observés) se démarquent bien par une variation d'entropie plus grande à leur frontière droite. Cependant, on observe aussi une zone relativement importante de confusion, qui confirme la nécessité de chercher la segmentation optimale d'une phrase, et non simplement les frontières de façon indépendantes les unes des autres.

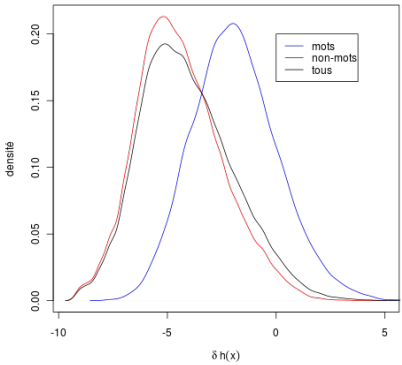

FIGURE 1 – Distribution des valeurs prises par la VBE à droite des bigrammes

On peut ensuite généraliser ce mode d'observation. En se demandant notamment si les valeurs de VBE à l'intérieur d'un mot sont aussi discriminantes que les valeurs qui en marquent les frontières. Pour différentes tailles de n-grammes mais en fixant n on va effectuer la même mesure et la même distinction entre mots et non-mots que précédemment mais cette fois-ci en prenant en compte les deux frontières gauche et droite ainsi que les valeurs observées à chaque inter-caractère à l'intérieur du n-gramme et dans chacun des deux sens de lecture possibles. Pour avoir une vue d'ensemble, on affiche ces résultats deux à deux sous la forme de courbes de niveaux. Les résultats pour les trigrammes sont présentés figure 2. On observe que des différentes valeurs observées les plus discriminantes sont sans conteste « gauche 1 » et « droite 3 », c'est à dire les entropies aux frontières. Il apparaît vraisemblable que la structure interne des unités morphologiquement complexes affectent la VBE, ce qui rend les valeurs internes plus difficiles à utiliser en pratique, contrairement à l'hypothèse suivie initialement dans (Magistry et Sagot, 2011).

4.2.2 Limites de la formulation par entropie

En présence de données aléatoires, on s'attend à ce que l'entropie de branchement diminue à mesure que la longueur de la chaîne considérée grandit. L'hypothèse de Harris nous fait dire que pour une chaîne donnée en langue naturelle, la variation de l'entropie de branchement lorsque l'on atteint une frontière sera anormalement élevée. Par ailleurs, on observe bien que pour des chaînes de même longueur, la variation de l'entropie de branchement aux frontières permet, au moins en partie, de distinguer les mots des non-mots. Toutefois, rien ne permet d'affirmer que cette distinction reste observable si l'on s'intéresse à des chaînes de longueurs différentes.

Nous avons donc cherché à observer les valeurs prises par la VBE aux frontières des n-grammes pour différentes valeurs de n. La figure 4.1.2 présente ces valeurs pour les uni- bi- et trigrammes. Elle montre qu'une normalisation ou qu'un recentrage de ces valeurs est nécessaire pour les rendre comparables.

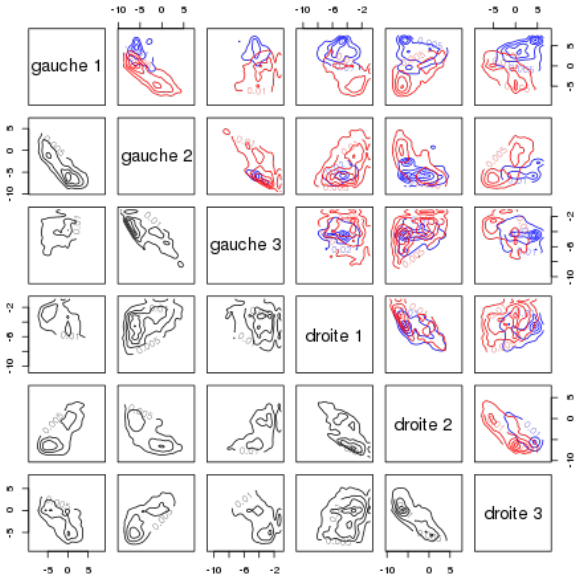

FIGURE 2 – Courbes de niveaux représentant la distribution des variations de l'entropie de branchement internes et aux frontières des trigrammes. La partie inférieur gauche du graphique (en noir) correspond à toutes les occurences de trigramme confondues, tandis que la partie supérieur droite distingue les mots (en bleu) des non-mots (en rouge). les dimensions indiquent qu'on s'intéresse à la variation d'entropie à gauche ou à droite d'un des trois caractères qui forment le trigramme, toujours en partant de l'extrémité opposée du trigramme.

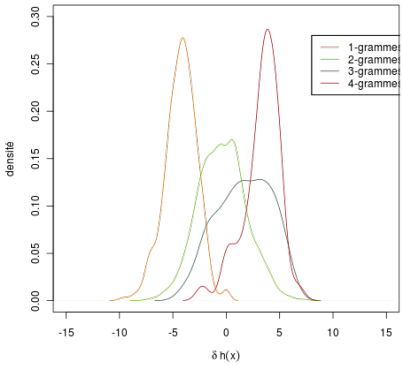

FIGURE 3 – VBE observée à droite des mots de différentes longueurs

4.2.3 Conséquences

Nous venons d'observer d'une part que l'information que l'on calcule aux frontières des mots est de loin la plus pertinente pour distinguer les unités lexicales recherchées, et d'autre part que si l'on cherche à comparer ces valeurs pour des unités de longueurs différentes, une normalisation est nécessaire. Ces observations motivent et permettent des modifications importantes du modèle de segmentation : les valeurs pertinentes sont moins nombreuses et indépendantes du contexte. Elles sont donc précalculables pour l'ensemble des n-grammes observés. De plus il nous faut les normaliser et chercher, pour une séquence de caractère donnée, à trouver la segmentation qui maximise ces mesures.

Ces modifications sont intégrées à notre algorithme de décodage présenté à la section suivante.

5 Algorithme de segmentation proposé

Les VBE ne sont pas directement comparables pour des chaînes de longeurs différentes, et doivent être normalisées. Nous recentrons les VBE, pour chaque longueur de chaîne, autour de 0. Pour cela, nous retranchons simplement à la VBE d'une chaîne de longeur k la moyenne des VBE de toutes chaînes de même longueur. Nous notons $\tilde{\delta h}_{\rightarrow}(x)$ et $\tilde{\delta h}_{\leftarrow}(x)$ les VBE normalisées. Pour simplifier, nous ne donnons que la définition de $\tilde{\delta h}_{\rightarrow}(x)$, celle de $\tilde{\delta h}_{\leftarrow}(x)$ en étant symétrique : pour chaque longueur k et chaque k-gramme x tel que $len(x) = k$, $\tilde{\delta h}_{\rightarrow}(x) = \delta h_{\rightarrow}(x) - \mu_{\rightarrow,k}$, où $\mu_{\rightarrow,k}$ est la moyenne des valeurs de $\delta h_{\rightarrow}(y)$ de tous les k-grammes y.

Il est important de noter que nous utilisons et normalisons la VBE et non l'entropie de branchement elle même. En effet, utiliser la BE contredirait l'hypothèse de Harris, puisque l'on ne

s'attendrait plus à ce que l'on ait $\tilde{h}(x_{0..n}) < \tilde{h}(x_{0..n-1})$ aux endroits qui ne sont pas des frontières de mots. De nombreux travaux utilisent pourtant la BE, normalisée ou non, et non la VBE, et obtiennent des résultats inférieurs à l'état de l'art (Cohen *et al.*, 2002).

Si nous ne basons nos décisions de segmentations que sur la VBE aux frontières de mots, chercher la meilleure segmentation d'une phrase revient à chercher en celle-ci les mots présentant les « meilleures frontières ». Cette qualité des frontières rejoint intuitivement (et empiriquement dans une certainement mesure, voir Magistry et Sagot (2011)) la notion d'autonomie syntaxique des unités qui composent la phrase. En termes de VBE, on peut définir la mesure d'autonomie d'un n-gramme comme $a(x) = \tilde{\delta}_{\leftarrow} h(x) + \tilde{\delta} h_{\rightarrow}(x)$.

On peut alors dire que plus l'autonomie $a(x)$ d'un n-gram x est grande, plus x est susceptible d'être un mot.

Avec cette mesure d'autonomie, on peut reformuler le problème de la segmentation d'une phrase comme la recherche du découpage qui maximise l'autonomie des mots qu'il délimite. Pour une séquence de caractères s, si on note $\mathrm{Seg}(s)$ l'ensemble de toutes les segmentations possibles, on cherche :

$$\arg\max_{W \in \mathrm{Seg}(s)} \sum_{w_i \in W} a(w_i) \times \mathrm{len}(w_i)$$

Où W est une segmentation délimitant les mots $w_0 w_1 \dots w_m$ et $\mathrm{len}(w_i)$ est la longueur d'un mot w_i, utilisée ici pour rendre comparables des segmentations aboutissant à des nombres de mots différents. Multiplier la mesure d'autonomie par la longueur du mot revient à attribuer un score aux caractères, qui contrairement aux mots sont en nombre constant entre les segmentations possibles d'une même chaîne.

Cette segmentation optimale en terme de VBEs est calculable simplement par programmation dynamique.

5.1 Décodage par programmation dynamique

Notre mesure d'autonomie d'un n-gramme donné est calculée à partir de tous ses contextes observés en corpus. Mais une fois calculée, elle ne dépend pas d'un contexte particulier. Elle ne dépend notamment pas du contexte observé spécifiquement au sein d'une chaîne en cours de segmentation.

Pour une chaîne donnée $u_{0..k}$ de longueur k, il y a 2^{k-1} segmentations possibles. Mais l'on peut remarquer que si l'on connaît la meilleure segmentation pour celle-ci et pour ses préfixes $u_{0..n}$, $n \leq k$, considérer un caractère supplémentaire et segmenter la chaîne $u_{0..k+1}$ ne nécessite que de considérer les appartenances possibles du caractère supplémentaire (le $k+1^{\text{ième}}$). Étant donné que nos mots sont contraints à être des séquences continues (insécables) de caractères, il nous suffit donc de considérer les cas suivants :

1. l'ajout du $k+1^{\text{ième}}$ caractère comme un mot de longueur 1 à fin de la meilleure segmentation de $u_{0..k}$
2. pour chaque prefixe $u_{0..n}$ de $u_{0..k}$ (avec $0 < n < k$), le cas où le $k+1^{\text{ième}}$ caractère est intégré à un mot unique de longueur $k-n$ qui vient s'ajouter à la fin de la meilleur segmentation de $u_{0..n}$

3. toute la chaîne $u_{0..k+1}$ est un mot unique de longueur $k+1$.

les cas 1 et 3 ci-dessus peuvent être vus comme les bornes du second cas, qui est le cas général si on prend $0 \leq n \leq k$. Ils sont explicités ici pour plus de clarté. Ce constat nous permet de reformuler la meilleure segmentation de $u_{0..k+1}$ à partir des meilleures segmentations de ses préfixes comme suit :

$$\underset{W \in \text{Seg}(u_{0..k+1})}{\arg\max} = \underset{V \in \bigcup_{n \leq k} \bigcup_{S \in \text{Seg}(u_{0..n})} S \cup \{u_{n..k+1}\}}{\arg\max} \sum_{w_i \in V} a(w_i) \times \text{len}(w_i)$$

Cette reformulation permet une programmation dynamique qui garde en mémoire les meilleures segmentations des $\text{Seg}(u_{0..n})$ et qui nous amène à ne considérer que $\sum_{n=2}^{k} n$ segmentations au lieu des 2^{k-1} théoriquement possibles. Cette méthode amène un surcoût négligeable pour $k < 5$ et devient de plus en plus intéressante à mesure que k grandit à partir de $k \geq 5$, ce qui est le plus souvent le cas.

6 Resultats et discussion

Nous avons évalué notre système sur les quatre corpus du *Bakeoff 2* et dans les configurations 2 et 3 telles que décrites à la Section 3. Nous comparons notre système (nVBE) aux résultats de Wang *et al.* (2011) ainsi qu'à notre propre implémentation de la stratégie « couper si une BE est croissante ». avec des variations de BE calculées dans les deux sens de lecture et pour toutes les longueurs de n-grammes, $1 \leq n \leq 6$. (à chaque position entre deux caractères, au plus 12 variations sont calculées, on segmente si au moins l'une d'entre elles est positive). Les résultats sont donnés Table 2. Les résultats filtrés par longueur de mot se trouvent Table 3.

Comme nous pouvons le voir, notre système est nettement meilleur que la stratégie de coupure sur accroissement de BE et obtient des scores comparables à ceux de ESA sans nécessiter de nombreuses itérations ni recourir à un paramètre.

Cela montre qu'on peut atteindre un bon niveau de segmentation en se basant uniquement sur une mesure de séparation. Lorsque celle-ci est maximisée pour une séquence donnée, il est raisonnable de penser qu'il existe une corrélation avec une éventuelle mesure de cohésion. Il n'est ainsi plus nécessaire d'avoir à trouver comment combiner les deux mesures.

On peut noter par ailleurs que l'évolution de nos résultats en fonction de la longueur des mots semble en accord avec la cohérence des guides d'annotation.

Nous ne pouvons fournir ici une analyse qualitative détaillée des résultats. Signalons tout de même que les erreurs observées nous semblent de même nature que nos observations antérieurs (Magistry et Sagot (2011)) et que celles présentes de la thèse de Jin. De nombreuses erreurs sont aussi liées aux dates et nombres écrits en chinois. Elles pourraient être écartées lors du prétraitement. D'autres erreurs concernent des morphèmes grammaticaux (« mots vides »)de haute fréquence et des affixes particulièrement productifs. Ces erreurs sont susceptibles de questionner les linguistes. Elle pourraient être corrigées en post-traitement par l'introduction de connaissances linguistiques.

Contrairement aux mots « pleins », ces mots vides ou morphèmes grammaticaux forment des classes fermées. De ce fait, introduire la connaissance linguistique nécessaire à leur bon traitement

System	AS	CITYU	PKU	MSR
Setting 3				
ESA bas	0.729	0.795	0.781	0.768
ESA haut	0.782	0.816	0.795	0.802
nVBE	0.758	0.775	0.781	0.798
Setting 4				
VBE> 0	0.63	0.640	0.703	0.713
ESA bas	0.732	0.809	0.784	0.784
ESA haut	0.786	0.829	0.800	0.818
nVBE	0.766	0.767	0.800	0.813

TABLE 2 – Évaluation sur les données du Bakeoff 2, suivant les configurations définies dans Wang *et al.* (2011). « Bas » et « haut » indiquent l'étendue des résultats obtenus par ESA pour différentes valeurs du paramètre du modèle. VBE> 0 segmente dès qu'une BE est croissante. nVBE correspond à la maximisation de la variation d'entropie de branchement normalisée aux frontières.

corpus	global	unigrammes	bigrammes	trigrammes
AS	0.766	0.741	0.828	0.494
CITYU	0.767	0.739	0.834	0.555
PKU	0.800	0.789	0.855	0.451
MSR	0.813	0.823	0.856	0.482

TABLE 3 – Resultats par longueur de mots (nVBE, configuration 4)

dans un système de segmentation ne nécessite qu'une quantité de travail limitée. Recourir à un système d'apprentissage supervisé ou symbolique pour traiter les classes de mots fermées et déléguer la gestion des classes ouvertes à un système non-supervisé nous semble être une voie prometteuse et linguistiquement pertinente.

Remarquons enfin que notre système obtient de bien meilleurs résultats sur les corpus de MSR et de PKU. Le corpus PKU étant le plus petit et AS le plus grand, la taille du corpus d'entraînement ne semble donc pas jouer à elle seule un rôle primordial pour expliquer les différences. En revanche, PKU est le corpus le plus homogène, il contient des articles qui sont tous issus du même journal. Le corpus AS au contraire est équilibré et présente une forte hétérogénéité des contenus. Le corpus CITYU est presque aussi petit que PKU mais contient des articles issus de journaux représentatifs de différentes variétés de mandarin, on peut donc s'attendre à ce que son contenu présente de grandes variations. Il semblerait donc que l'homogénéité des données d'entraînement soit aussi importante sinon plus que la quantité des données utilisées pour le bon fonctionnement du système présenté ici. Cette observation devra être vérifiée dans de prochains travaux. Si elle se confirmait, une étape de classification automatique des données d'entraînement pourrait être un prétraitement essentiel.

Références

COHEN, P., HEERINGA, B. et ADAMS, N. (2002). An unsupervised algorithm for segmenting categorical timeseries into episodes. *Pattern Detection and Discovery*, pages 117–133.

EMERSON, T. (2005). The second international chinese word segmentation bakeoff. *In Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, volume 133.

FENG, H., CHEN, K., DENG, X. et ZHENG, W. (2004). Accessor variety criteria for chinese word extraction. *Computational Linguistics*, 30(1):75–93.

GOLDWATER, S., GRIFFITHS, T. et JOHNSON, M. (2006). Contextual dependencies in unsupervised word segmentation. *In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 673–680.

HARRIS, Z. S. (1955). From phoneme to morpheme. *Language*, 31(2):190–222.

HUANG, C. et ZHAO, H. (2007). 中文分词十年回顾(chinese word segmentation : A decade review). *Journal of Chinese Information Processing*, 21(3):8–20.

JIN, Z. et TANAKA-ISHII, K. (2006). Unsupervised segmentation of chinese text by use of branching entropy. *In Proceedings of the COLING/ACL on Main conference poster sessions*, pages 428–435.

KEMPE, A. (1999). Experiments in unsupervised entropy-based corpus segmentation. *In Workshop of EACL in Computational Natural Language Learning*, page 713.

MAGISTRY, P. et SAGOT, B. (2011). Segmentation et induction de lexique non-supervisées du mandarin. *In Actes de TALN 2011, Montpellier*, pages 333–344.

MOCHIHASHI, D., YAMADA, T. et UEDA, N. (2009). Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling. *In 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 : Volume 1-Volume 1*, pages 100–108.

SPROAT, R. W. et SHIH, C. (1990). A statistical method for finding word boundaries in chinese text. *Computer Processing of Chinese and Oriental Languages*, 4(4):336–351.

TANAKA-ISHII, K. (2005). Entropy as an indicator of context boundaries : An experiment using a web search engine. *In International Joint Conference on Natural Language Processing (IJCNLP-2005)*, pages 93–105.

T'SOU, B., LIN, H., LIU, G., CHAN, T., HU, J., CHEW, C. et TSE, J. (1997). A synchronous chinese language corpus from different speech communities : Construction and applications. *Computational Linguistics and Chinese Language Processing*, 2(1):91–104.

WANG, H., ZHU, J., TANG, S. et FAN, X. (2011). A new unsupervised approach to word segmentation. *Computational Linguistics*, 37(3):421–454.

ZHANG, Y. et CLARK, S. (2010). A fast decoder for joint word segmentation and POS-tagging using a single discriminative model. *In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 843–852.

ZHAO, H. et KIT, C. (2008). An empirical comparison of goodness measures for unsupervised chinese word segmentation with a unified framework. *In The Third International Joint Conference on Natural Language Processing (IJCNLP-2008), Hyderabad, India*.