Segmentation automatique des mots en chinois

Introduction

En l'absence de délimiteurs de mots, la tâche de segmentation est la première étape du TAL en chinois.

Il existe plusieurs formes d'ambiguïté, lexicales ou syntaxiques, donnant lieu à des variations de signification de mots ou de configurations structurelles. En TAL appliqué au chinois, les ambiguïtés peuvent être liées aux caractéristiques des sinogrammes (par exemple les homographes) et

se réduit à la désambiguisation lexicale.

L'ambiguïté structurelle en revanche résulte de la possibilité de pouvoir analyser de plusieurs manières valides une unité linguistique complexe. Ce problème correspond à celui

de l'analyse syntaxique et à la segmentation, car dans ce cas plusieurs segmentation ou annotations sont possibles.

Cette ambiguïté peut prendre la forme d'un chevauchement:

美国会 peut se comprendre comme

- 美国会 congrès américain
- ou 美国+会 les USA peuvent

Elle peut prendre la forme d'une ambiguïté combinatoire:

才能 peut se comprendre comme un groupe nominal composé ou comme la combinaison de l'adverbe 才 (seulement) et du verbe 能 (pouvoir).

Ce type d'ambiguïté est la plus courante car elle résulte de la propension de formes verbales ou nominales en chinois ancien à devenir des adverbes en chinois moderne, d'où l'expression citée souvent en syntaxe diachronique du chinois:"la syntaxe d'aujourd'hui est le lexique de demain".

Les formes d'ambiguïté peuvent bien entendu se conjuguer en augmentant la complexité: 太平淡 peut être analysé comme 太平淡 (très fade), 太平 (pacifique) 太(trop) 平 (plat) 淡 (brut).

Les manières de lever cette ambiguïté dans la cas d'une analyse contextuelle, par exemple si une analyse d'une phrase complète ne comprends aucun verbe, elle est éliminé pour une autre solution faisant apparaître le verbe.

Mais elle est impossible par exemple lorsqu'un groupe verbal peut-être décomposé entre plusieurs verbes et adverbes différents:

应用+于 (appliquer à) 应+用于 (devoir utiliser).

Un cas équivalent et non résolvable par des règles est le cas particulier des sigles, les noms propres ou les translittérations, qui sont l'abréviation d'un syntagme long en un ensemble de composantes lexicales qui pourraient être indépendantes.

Les caractère chinois peuvent donc apparaître dans des positions différentes dans des mots différents, et l'absence de délimiteur amplifie la difficulté de l'analyse. Il n'est pas possible d'isoler des liste de sinogrammes mutuellement exclusifs, qui auraient des distributions distinctes. Les caractères eux mêmes ne peuvent peuvent pas être utilisés comme des indicateurs de fin ou de début de mot:

en fin de mot => 生产(produire)

en position indépendante => 产+小麦(cultiver du blé)

en milieu de mot => 生产线(ligne de production)

en début de mot =>产生 (inventer)

Comme pour des langues comme le finnois, il est impossible dans une application TAL de lister exhaustivement toutes les formes lexicales qui peuvent être rencontrés, et cela bien que le nombre de caractères chinois reste constant.

Les néologismes sont aisément produits selon différents mécanismes (combinaison phonétique ou sémantiques de caractères dans des syntagmes complexes).

Algorithmes d'analyse

Approche basée sur les sinogrammes

Cette approche peut prendre en compte les caractère individuellement ou les groupes de caractères.

Le fait de considérer les caractères individuellement est efficace dans les textes où la majorité des caractères sont indépendants, ce qui correspond essentiellement au chinois classique et archaïque, ou certaines expressions formelles/idiomatiques.

Le chinois moderne utilisant à 95% des mots constitués de 2 ou plusieurs caractères, la majorité des textes gagnera a être analysée par groupes de caractères.

Entropie maximale

La méthode classiquement utilisée est celle de l'entropie maximale: chaque caractère est se voit attribuer statistiquement une classe grâce à son contexte. On observe un contexte linguistique du caractère ($b \in B$) pour prédire l'appartenance du caractère à une classe ($a \in A$). La procédure consiste à construire un classificateur

$\alpha : B \rightarrow A$ implémenté à partir une probabilité conditionnelle p telle que $p(a|b)$ est la probabilité de la classe a dans le contexte b.

Pour définir le contexte, on associe chaque caractère à sa position dans le texte. Pour chaque mot, un caractère va apparaître à une position fixe.

Classer les caractère en différentes classes consiste à décider pour chaque caractère s'il apparaît en début, milieu, fin de mot ou si le caractère est isolé.

par exemple pour 本, les positions sont:

一 本 书 indépendant

剧本 fin

本来 début

基本上 milieu

Cette information sera représentée par une étiquette pour le classificateur.

les traits linguistiques considérés dans la segmentation sont les positions possibles d'un

caractère dans un mot:

(a) défaut

(b) caractère courant (C0)

(c) précédent (C−2 , C−1 , C1 , C2)

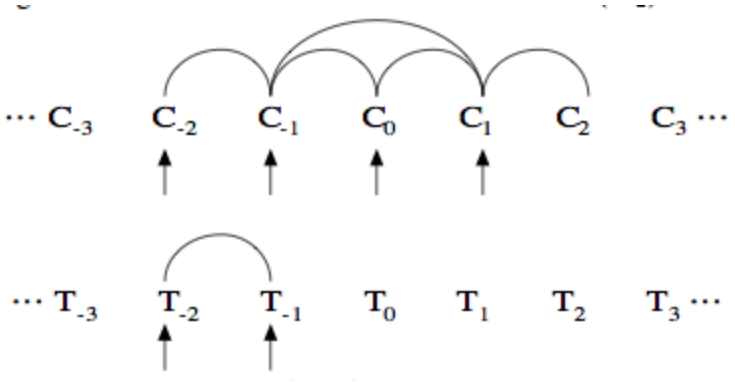
(d) prochain (C-1 C0, C0 C1),

deux prochains (C-2 C-1), et deux précédents (C1 C2)

(e) prochain et précédent (C-1 C1)

(f) étiquette du précédent caractère (T-1), et étiquette du caractère deux positions avant

(T-2)



On établis ensuite des relations de co-occurrences :

爱 sera étiqueté comme étant en début de mot si le prochain caractère est 护. mais si le

caractère précédent est 热, alors il sera étiqueté comme étant en fin de mot.

La probabilité associant un contexte h à une étiquette t est définie par:

$$p(h,t) = \pi\mu\prod_{j=1}^{k}\alpha_j^{f_j(h,t)}$$

où π est une constante de normalisation, {μ,α1,...,αk} sont paramètres du modèle et

{f1,...,fk} sont les traits définis tels que fj(h,t)∈{0,1}.

Chaque trait fj correspond à un paramètre α j, qui permet de pondérer ce trait. Pendant l'apprentissage, pour une séquence de caractères {c1,...,cn} et leurs étiquettes {t1,...,tn} ,on cherche à déterminer les paramètres {μ,α1,...,αk} qui obtiennent la vraisemblance la plus élevée sur les données à partir de p:

$$L(P) = \prod_{i=1}^{n} P(h_i, t_i) = \prod_{i=1}^{n} \pi\mu \prod_{j=1}^{k} \alpha_j^{f_j(h_i, t_i)}$$

Pour couple (h,t) donné, un trait linguistique doit être assez bien paramétré pour permettre de prédire t, et la définition des traits linguistique est le principal facteur de réussite de cette méthode.

champs conditionnels aléatoires

Les publications plus récentes tendent à remplacer l'entropie maximale par les champs conditionnels aléatoires (CRF) qui est une représentation probabiliste de la structure d'un graphe. Le graphe en question est la chaîne linéaire des caractères, qui correspond à une machine à état finis.

Les paramètres

$\lambda=\{\lambda 1,...,\lambda k\}$

Définissent une probabilité conditionnelle pour la séquence d'étiquettes

$Y=y1,...,yT$

sur la séquence entrée

$X=x1,...xT$

Les CRF sont entraînés en utilisant l'estimation maximale de probabilité. C'est un modèle discriminant qui permet d'identifier des traits corrélés des chaînes de caractères. Ce qui permet une utilisation dans de nombreux domaines du TAL nécessitant un étiquetage.

On représente une chaîne de caractère x et une séquence y d'étiquette associées par la formule:

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp \sum_{t=1}^{T} \sum_{k=1}^{K} \lambda_k f_k(t, y_t, y_{t-1}, \mathbf{x})$$

où fk (t, yt , yt−1 , x) est une fonction binaire indiquant la présence du trait linguistique k, λk est la pondération de ce trait, et Z(X) est la fonction de normalisation définie par:

$$Z(\mathbf{x}) = \sum_{y} \exp \sum_{t=1}^{T} \sum_{k=1}^{K} \lambda_k f_k(t, y_t, y_{t-1}, \mathbf{x}).$$

elle assure que la somme des probabilités de toutes les séquences de la chaîne soit égale à

1.

$$\mathcal{L}_\lambda = \sum_{i=1}^{N} \log p(\mathbf{y}^{(i)}|\mathbf{x}^{(i)}) + R(\lambda),$$

la pondération des traits est déterminée de manière à obtenir la log-vraissemblance

maximale des données d'apprentissage.

$$R(\lambda) = \frac{1}{C} \sum_{k=1}^{K} |\lambda_k|,$$

R(λ) est un nivellement qui assure que le modèle ne dépasse pas les limites des données

d'entraînement.

Le degré de nivellement est défini par une constante C.

Représentation des niveaux de segmentation:

s représente les symboles terminaux (traits linguistiques) indexés par leurs positions.

h représente les mots.

Symboles:

s−2 , s−1 , s0 , s+1 , s+2

Unigrammes

s−2 s−1 , s−1 s0 , s0 s+1 , s+1 s+2

Bigrammes

s−3 s−2 s−1 , s−2 s−1 s0 , s−1 s0 s+1 , s0 s+1 s+2 , s+1 s+2 s+3

Trigrammes

Mots:

h−2 , h−1 , h0 , h+1 , h+2

Unigrammes

h−2 h−1 , h−1 h0 , h0 h+1 , h+1 h+2

Bigrammes

h−1 h0 h+1

Trigrammes

Les CRF permettent d'utiliser comme paramètre de nombreux traits linguistiques corrélés.

Étant entraînés de manière discriminante, ils sont plus performants que les modèle

génératifs utilisant les même traits.

Approche basée sur les mots.

Les approches basées sur les mots visent à extraire des mots complets dès la première

analyse.

Méthode par dictionnaire

La méthode utilisée est celle de la plus longue correspondance et peut prendre deux

formes:

Analyse frontale (Forward matching method)

Il s'agit d'une analyse de type gloutonne qui cherche associer la plus longue chaîne de caractères possible à partir de chaque caractère de la phrase correspondant à une entrée de dictionnaire. Si aucune correspondance n'est trouvée, le caractère est isolé et l'analyse reprend sur le caractère suivant jusqu'à la fin de la phrase.

Analyse arrière (backward matching method).

Il s'agit d'un optimisation de la méthode précédente. On part du dernier caractère de la phrase en utilisant un dictionnaire inversé. Cette méthode a montré en général un meilleur taux de fiabilité, en particulier pour différencier les particules grammaticales isolées des

mots complexes composé de ces particules:鞋子 / 和服 / 装 鞋子 / 和/服装 par FMM 和服

=>kimono par BMM 和=> conjonction de coordination.

Méthode statistiques

Information mutuelle des bigrammes

On s base sur le taux d'information mutuel de deux caractères adjacents pour décider s'ils forment un mot de deux caractères.

L'information mutuelle des caractères adjacents x et y est définie par:

$I(x,y) = p(x,y)/p(x)p(y)$

avec p(x,y) la probabilité que x et y soient adjacents dans le corpus et p(x), p(y) celles qu'ils s'y trouvent isolés.

La paire de caractère adjacents avec le taux d'information mutuelle le pus élevé au delà d'un seuil déterminé est regroupé dans un mot.

L'analyse est continuée pour tous les caractères restants jusqu'à ce qu'aucun bigramme n'ait un taux d'information mutuelle supérieure au seuil.

Pour les mots composés de plusieurs caractères, l'algorithme d'espérance-maximisation est utilisé.

Dans ce modèle les mots peuvent être de longueur allant de 1 à n et ont chacun une probabilité d'occurrence inconnue.

Les mots sont indépendants les uns des autres.

Les mots sont des n-grammes candidats.

Les probabilités des chaque mots sont d'abord assignées de manière aléatoires et utilisées

pour segmenter le texte. Les probabilités des mots sont ensuite réévaluées à partir des

résultats de la segmentation, et on resegmente en fonction des nouvelles probabilités.

L'algorithme est répété jusqu'à convergence.

Méthodes statistiques utilisant un dictionnaire

Un dictionnaire est représenté par un transducteur pondéré à état fini.

Chaque mot est une séquence de transitions représentant association entre un les

caractères du mot et leur prononciations.

Les mots sont pondérés en fonction de leur coût. Ce coût est calculé en utilisant les log-probabilités négatives dérivées de la fréquence du mot dans un corpus.

La segmentation produite par le transducteur est celle sont le coût est le moins élevé.

Le coût pour un mot est évalué par la formule c(m):

T est la taille du corpus.

c(m) = -log(fréquence(m)/T)

Cette approche a été utilisée pour les mots dérivés et les entités nommées. Elle nécessite un traitement particulier pour les mots inconnus.

traiter les mots inconnus

La limite de cette approche est la nécessité de trouver un moyen d'évaluer la probabilité des différents types de mots inconnus.

Une approche utilisée est de différencier le lexique connu (basé sur un dictionnaire) et le lexique candidat (contenant tous les n-grammes hors du dictionnaire). On utilise l'algorithme d'espérance-maximisation sur un corpus d'apprentissage pour suggérer les n-grammes candidats à ajouter au lexique connu. Une fois ces mots ajoutés, on réinitialise l'algorithme en assignant la moitié de la probabilité totale à l'ensemble du lexique connu, ce qui a pour effet de guider la segmentation par cette classe du lexique.

Une fois que l'algorithme d'espérance-maximisation est stabilisé, on utilise l'information mutuelle pour éliminer les n-grammes les plus longs au profit des primitives les plus courtes car cet algorithme a tendance à pénaliser les segmentations comprenant le plus de parties.

Méthodes utilisant une grammaire.

Les méthodes précédentes ne font appel à aucune connaissance linguistique sur les phrases analysées. Chaque phrase est considérée indépendemment du texte la contenant. Cependant les éléments syntaxique peuvent être utilisés pour différencier les morphèmes et mots monosyllabiques.

Les mots monosyllabique sont considéré comme des unité lexicales et les caractères composant de mots inconnus et ne figurant jamais comme mot monosyllabique sont considérés comme des unités non lexicales. Les contextes des instances de ces unités lexicales et non lexicales dans le texte sont dérivés en règles d'une grammaire contextuelle. Ces règles sont hiérarchisées en fonction de leu précision à identifier les caractères isolés, et celles dont la précision dépasse un seuil défini sont utilisées séquentiellement pour

distinguer les caractères isolés des caractère strictement utilisés comme composant lexicaux.

Il est possible d'améliorer la précision en soumettant les mots inconnus à une grammaire hors-contexte pour analyser leur morphologie, par exemple:

1. Mot-Inconnu -> monosyllabique + n-gramme

2. Mot-Inconnu -> n-gramme

3 .Mot-Inconnu -> Mot-Inconnu + n-gramme

Ces règles sont conjuguées avec des contraintes linguistiques (sur les catégories syntaxiques d'un symbole) ou statistiques (sur la fréquence minimale d'un symbole) et hiérarchisées en fonction de la fréquence des symboles en partie droite des règles.

On peut enfin adjoindre des règles syntaxiques en utilisant un parseur pour apporter des informations sur la structure de la phrase.

Approches par apprentissage

Apprentissage par transformation

L'apprentissage par transformation utilise un un corpus de référence pré-segmenté et un segmenteur de départ qui peut être naïf (chaque caractère est traité comme un mot) ou sophistiqué (utilisant un méthode de maximum matching).

On compare la segmentation de référence à la segmentation du même texte par l'algorithme de départ pour identifier les règles qui permettent d'obtenir le meilleur gain de

précision. Ces règles sont ajoutées au segmenteur de départ et on réitère l'apprentissage

jusqu'à ce que le gain de précision soit en dessous d'un seuil déterminé.

Les règles sont hiérarchisées en fonction du gain associé et le segmenteur est utilisé sur de

nouveaux textes.

Modèle de Markov caché.

Cette approche utilise l'étiquetage syntaxique.

Soit S une séquence de caractère, et S(M) la séquence composant le mot M. Les étiqueter

morpho-syntaxiques forment une séquence E= e1,...,en pour une segmentation

M=m1,...,mn donnée. La segmentation retenue est celle qui correspond à la probabilité

suivante:

$$M,E = \arg \max_{M,E,M(S)=S} P(E,M|S)$$

$$= \arg \max_{M,E,M(S)=S} P(M,E)$$

$$= \arg \max_{M,E,M(S)=S} P(M|E)P(E)$$

La probabilité de l'étiquetage P(E) est déterminée par l'étiquette précédente et la probabilité conditionnelle P(M|E) est déterminée par l'étiquetage du mot. Le MMC suppose que chaque mot engendre à un état caché qui correspond à l'étiquette morpho-syntaxique de ce mot.

Une étiquette ei-1 est suivie par une étiquette ei avec la probabilité P(ei|ei-1) et engendre un mot avec la probabilité P(mi|ei).

Une approximation de ces deux probabilités peut être réécrite par:

$$P(M|E) \triangleq \prod_{i=1}^{n} P(m1|e1)$$

$$P(E) = \prod_{i=1}^{n} P(ti|ti-1)$$

Les probabilités sont estimées à partir de la fréquences des occurrences d'un corpus étiqueté en utilisant le maximum de vraisemblance.

F(X) est la fréquence des occurrences dans le corpus étiqueté, <mi,ei> sont les co-occurrences d'un mot et d'une étiquette, et <ei,ei-1> les co-ocurrences de deux étiquettes.

$P(mi|ei) = F(<mi,ei>)/F(ei)$

$P(ei|ei-1) = F(<ei,ei-1>)/F(ei-1)$

La segmentation possible d'une phrase peut être représentée par un treillis dont les noeuds correspondent à un mot possible avec son étiquette associée. La séquence de mots la plus probable est calculée par l'algorithme de Viterbi qui consiste à sélectionner la séquence d'étiquettes la plus probable pour la séquence de caractère donnés en comparant les différents chemins possibles sur le graphe.

Cette approche ne permet pas de segmenter les mots inconnus, pour lesquels il faut faire appel à un traitement séparé.

Désambiguïsation de la segmentation

Dictionnaire

La correspondance maximale est la méthode basée sur un dictionnaire la plus simple, en comparant les correspondances avant et arrières, on peut déterminer les positions dans lesquelles les ambiguïtés peuvent exister.

pour le syntagme 即将来临时 on peut obtenir les segmentations:

Par analyse frontale: 即将/来临/时

Par analyse arrière: 即/将来/临时

On compare les deux analyses A1 et A2. Les mots inconnus sont segmentés en caractères uniques si les deux analyses sont correctes.

L'ambiguïté est résolue par une classification binaire:

Pour la chaîne de caractère ambiguë la plus longue et son contexte C={$m_{-a}...m_{-1},m_1...m_b$}, le score calculé par la fonction G(Seg,C)

où Seg ? {A1,A2} est défini par:

$$G=p(Seg)\prod_{i=a...-1,1...b} p(m_i|Seg)$$

l'ambiguïté est tranché par la décision binaire:

si G(A1,C) > G(A2,C) on choisit A1

si G(A1,C) <G(A2,C) on choisit A2

-Si A1=A2, les segmentation sont équivalentes et on peut choisir l'une des deux.

-Sinon on choisit celle à laquelle la fonction G attribue le meilleur score.

La méthode par dictionnaire repose sur un algorithme glouton et sa précision dépends de la taille du dictionnaire. Il est impossible de détecter des mots inconnus car seuls les mots présents dans le dicitonnaire peuvent être segmentés de manière fiable.

Méthode statistiques

Le modèle d'espace vectoriel permet de modéliser le contexte des mots ambigues.

L'ambiguité de la segmentation est traitée comme une ambiguité de sens.

Dans ce modèles, tous les mots co-occurents d'un mot ambigue peuvent être extraits sous la forme d'un vecteur représentant le contexte. Ce contexte est souvent réduit à ±3 mots autour mot candidat considéré.

Les six mots pris dans ce contexte sont divisés en 4 régions:

R1 : mot-3 et mot-2

R2 : mot-1

R3 : mot+1

R4 : mot+2 et mot+3

exemple de Xiao, 2001:

Fig. 1: The window of context of *w*

La forme segmentée possible i d'un mot w est i=1 quand w est composant d'un mot et i=2 quand w est un mot isolé.

Pour calculer le poids du mot tj dans la région Rk pour la fome segmentée i de w, la forme générale est

(i=1,2;j=1;=1...4) et est définie par:

Di est l'ensemble des phrases contenant w dans i (le corpus d'entrainement de w)

n est le nombre de mots dans l'union de D1 et D2

tf_{ijk} fréquence du mot t_j dans la région R_k de l'ensemble D_i.

tf_{qjk} fréquence du mot t_j dans la région R_k de la phrase Q contenant le mot w.

df_{jk} le nombre des collections (D_1 et D_2) contenant t_j dans la région R_k. Prends une valeur de 0 à 2.

$$idf_{jk} = \log(d/df_{jk})$$

La forme classique de TF et IDF est

$$TF_{ijk} = tf_{ijk}$$

$$IDF_{jk} = idf_{jk} = \log(d/df_{jk})$$

Le coefficient de similarité entre w dans la phrase Q et la segmentation i de w est définie par:

$$CS(Q,i)=\sum_{j=1}^{n}\sum_{k=1}^{4} tf_{qij} \times d_{ijk}$$

Le mot ambigue w dans Q prendra la forme segmentée dont le coefficient de similarité avec i sera le plus élevé.

Le modèle vectoriel est prends dans le cas de chinois une dimension importante en raison du lexique vaste, et les données sont donc dispersées.

Pour résoudre ce problème, les les mots dont la fréquence est faible sont remplacés par un code sémantique extrait d'un thésaurus, ce qui permet de généraliser le modèle.

Les codes sémantiqes sont déterminés par des règles syntaxique et des annotations sémantiques en utilisant un thésaurus comme 同义词词林.

Conclusion

Plusieurs université e centre de recherches proposent des standards de segmentation sous la forme de règles propres et de corpus de références.

L'université de Pékin qui se base principalement sur la définition exhaustive du lexique et des règles de segmentation, l'Academia Sinica qui propse des règles plus simples, L'université de Pennsylvanie qui a developpé un standard proche de celui de l'unversité de Pékin utilisé pour le Chinese Treebank.

L'évaluation d'un segmentaiton utlise les mesures habituelles de rappel précision et F-mesure mais en faisant la distinction entre le rappel pour la segmentation des mots inconnus et des mots connus.

La tâche de segmentation fait de plus en plus appel à des traitements annexes pour améliorer sa précision, notament la reconnaissance d'entités nommées et l'analyse des structues syntaxiques.

Bibliographie

Unknown word detection for Chinese by a corpus-based learning methodKJ Chen, MH Bai - Computational Linguistics, 1998 - 140.109.19.102

Unigram language model for Chinese word segmentationA Chen, Y Zhou, A Zhang, G Sun - SIGHAN Workshop on Chinese Language Processing , 2005

Word association norms, mutual information, and lexicographyKW Church, P Hanks - Computational linguistics, 1990 - portal.acm.org

Chinese word segmentation and named entity recognition: A pragmatic approachJ Gao, M Li, A Wu, CN Huang - Computational Linguistics, 2005 - portal.acm.org

Discovering Chinese words from unsegmented text (poster

abstract)X Ge, W Pratt, P Smyth - 22nd annual international ACM SIGIR Conference, 1999 - portal.acm.org

Error-driven learning of Chinese word segmentationJ Hockenmaier, C Brew - Language, 1998 - Citeseer

Unsupervised training for overlapping ambiguity resolution in Chinese word segmentation

M Li, J Gao, C Huang, J Li - International Journal of COmputational Linguistics and Chinese Language Processing, 2003 - portal.acm.org

Segmentation standard for Chinese natural language processingCR Huang, KJ Chen, LL Chang - International Journal of COmputational Linguistics and Chinese Language Processing, 1996 - portal.acm.org

Conditional random fields: Probabilistic models for segmenting and labeling sequence dataJ Lafferty, A McCallum, F Pereira - Proceedings of the International Conference on Machine Learning , 2001 - Citeseer

Unsupervised training for overlapping ambiguity resolution in Chinese word segmentationM Li, J Gao, C Huang, J Li - second SIGHAN workshop on Chinese Language Processing, 2003 - portal.acm.org

A Unicode based adaptive segmentorQ Lu, ST Chan, RF Xu, TS Chiu, BL Li, - Journal of Chinese Langage and Computing , 2003 - portal.acm.org

Covering ambiguity resolution in Chinese word segmentation based on contextual informationX Luo, M Sun, BK Tsou -

Proceedings of COLING , 2002 - portal.acm.org

Unknown word detection and segmentation of Chinese using statistical and heuristic knowledgeJY Nie, ML Hannan, W Jin - Communications of COLIPS, 1995 - iro.umontreal.ca

Chinese word segmentation and information retrievalD Palmer, J

Burger - AAAI Spring Symposium on Cross-Language Text

Mingin, 1997 - aaai.org

Chinese segmentation and new word detection using conditional

random fieldsF Peng, F Feng, A McCallum - Proceedings of the

20th international conference on Chinese Language Processing,

2004 - portal.acm.org

The first international Chinese word segmentation bakeoffR

Sproat, T Emerson - second SIGHAN workshop on Chinese, 2003

- portal.acm.org

statistical method for finding word boundaries in Chinese textR

Sproat, C Shih - Computer Processing of Chinese and Oriental

Languages, 1990

A stochastic finite-state word-segmentation algorithm for ChineseR Sproat, W Gale, C Shih, N Chang - Computational linguistics, 1996 - portal.acm.org

A compression-based algorithm for Chinese word segmentationWJ Teahan, R McNab, Y Wen, IH - Computational Linguistics, 2000 - portal.acm.org