SibyMot : Modélisation stochastique du langage intégrant la notion de *chunks*[1]

Igor Schadle, Jean-Yves Antoine, Brigitte Le Pévédic, Franck Poirier
Laboratoire VALORIA, Université de Bretagne Sud (EA 2593)
(igor.schadle@univ-ubs.fr)

Résumé – Abstract

Cet article présente le modèle de langage développé pour le système Sibylle, un système d'aide à la communication pour les personnes handicapées. L'utilisation d'un modèle de langage permet d'améliorer la pertinence des mots proposés en tenant compte du contexte gauche de la saisie en cours. L'originalité de notre modèle se situe dans l'intégration de la notion de *chunks* afin d'élargir la taille du contexte pris en compte pour l'estimation de la probabilité d'apparition des mots.

We present in this article the language model of Sibyl, a new Alternative and Augmentative Communication (AAC) system. The use of language modeling improves the relevance of displayed words by taking into account the left context of the current sentence. The originality of our model is to introduce chunking. This enlarges the context taken into account to estimate the words probability.

Mots Clés – Keywords

Aide à la communication, modélisation stochastique du langage, n-gramme, *chunks*.

AAC, stochastic language modeling, n-gram, chunks.

1 Handicap et communication

On appelle système d'aide à la communication tout système visant à suppléer ou restaurer, ne serait-ce que partiellement, la fonction de communication d'une personne handicapée. Dans le cas d'un handicap physique lourd (troubles de la parole et facultés motrices réduites), les modalités de communication sont limitées. Sur ordinateur, une solution consiste à composer les messages à l'aide d'un clavier simulé (clavier présenté à l'écran) via une interface adaptée. Lorsque l'interface d'accès n'autorise que l'équivalent du simple clic, comme le bouton poussoir, la saisie sur clavier simulé est réalisée par un système de défilement automatique.

[1] Activités de recherche financées par le Conseil Régional de Bretagne

Un curseur met en évidence les lettres une à une, à intervalle régulier, et l'utilisateur n'a plus qu'à valider lorsque le curseur pointe sur la lettre désirée. L'inconvénient majeur de ces aides est l'extrême lenteur d'écriture. Là où la communication orale permet un débit de l'ordre de 150 mots à la minute, ces aides ne permettent qu'une écriture autour de 5 mots à la minute. Pour accroître cette vitesse, le système Sibylle, développé au laboratoire VALORIA de l'Université de Bretagne Sud, propose deux aides complémentaires. La première, SibyLettre est un système de prédiction de lettre qui permet une sélection plus rapide des lettres (Schadle et al., 2001). La deuxième, SibyMot, est un système de prédiction de mot. Le système affiche une liste de mots, mots considérés comme les plus probables en fonction du contexte gauche de la phrase. En sélectionnant les mots dans la liste, l'utilisateur évite leur saisie complète. Comme d'autres systèmes de communication issus de la recherche, HandiAS (Maurel, Le Pévédic, 2001) ou VITIPI (Boissière, 2000), le système Sibylle utilise un modèle de langage avancé pour établir une liste de mots pertinente. Ce modèle est le sujet de ce présent article.

2 Idées et principes

2.1 Modélisation n-gramme

Le point de départ de notre modèle est le modèle statistique n-gramme gramme utilisé dans le cadre de la modélisation probabiliste du langage et issu de la théorie de l'information (Jelinek, 1976). Dans l'objectif de prédire des mots, le problème principal de ce modèle est de ne tenir compte que des derniers mots. De nombreuses variations ont été proposées pour améliorer l'utilisation de ce contexte. Cependant, malgré ces améliorations, le contexte reste à très courte distance, de l'ordre de deux à trois mots. Pour accroître la taille du contexte et capter de manière plus efficace les dépendances à plus longue distance, le modèle présenté, et c'est son originalité, propose d'intégrer la notion de *chunks*.

2.2 Analyse en *chunks*

L'analyse en *chunks* consiste à décomposer une phrase en syntagmes minimaux non récursifs. Par rapport à l'analyse syntaxique, l'analyse en *chunks* se distingue par le fait qu'elle ne cherche ni à donner les fonctions syntaxiques des syntagmes, ni à en établir les dépendances. (Abney, 1991) a utilisé les *chunks* comme étape préliminaire à l'analyse syntaxique. Depuis, la notion de *chunks* a largement été réutilisée en linguistique informatique : pour la reconnaissance de la parole, l'analyse syntaxique, la compréhension de la parole, etc. Dans le cadre de notre modèle, nous utilisons les *chunks* pour l'estimation du mot à venir. Ici, l'intérêt de cette analyse est de structurer le contexte gauche du mot à prédire. En particulier, avec les têtes des chunks (leur mot principal), elle permet de mettre en avant des mots pertinents pour la prédiction. Ainsi, au contexte des *n-1* derniers mots, nous associons un contexte des *n-1* dernières têtes de *chunks*. Sur l'exemple du début de phrase : « [*l'année**] [*du dragon**] [*a ... (commencé)* », le contexte tri-gramme pour le mot « commencé » est « dragon a », tandis que le contexte considérant les deux dernières têtes de *chunks* est « année dragon ». Ce dernier fait apparaître le mot « année », un bon *prédicteur* pour le mot « commencé ». Les *chunks* permettent donc de capter des dépendances à plus longue distance que le modèle n-gramme, tout en restant dans le cadre de la modélisation robuste du langage.

2.3 Les lemmes

Une autre des difficultés du modèle n-gramme est liée à l'espace des paramètres à estimer. Si V est la taille du vocabulaire, alors le nombre de paramètres est de l'ordre de V^N. Relativement à l'anglais, ce problème est accru en français par sa richesse flexionnelle. (Cerf Danon, El-Bèze, 1991) donnent un rapport formes fléchies/lemme de 2 en anglais contre 7 en français. Nous avons donc privilégié le lemme comme unité lexicale, la probabilité d'une forme fléchie étant exprimée comme la probabilité combinée du lemme et de la flexion.

3 Modélisation

Après avoir exposé les idées principales de notre modèle, nous allons maintenant décrire de manière plus détaillée son fonctionnement. Pour permettre une prédiction fondée sur les *chunks*, SibyMot est composé de deux modules : un *analyseur* chargé de construire une représentation de la phrase en *chunks* et un *prédicteur* qui délivre la probabilité d'apparition des mots du lexique. Rappelons que dans le cadre de l'application Sibylle, le modèle est utilisé en mot à mot. Les deux étapes analyse et prédiction sont indépendantes et, en particulier, l'analyseur peut être utilisé seul pour des tâches d'étiquetage et de segmentation.

3.1 Partie analyse

En ce qui concerne l'analyseur, la segmentation d'un énoncé en *chunks* correspond en TAL à une *analyse de surface* (*shallow parsing*). Dans notre système, l'analyse est chargée de déterminer pour chaque mot son lemme et son étiquette grammaticale. De plus, au niveau du *chunk*, elle délivre l'étiquette associée au *chunk* (sa catégorie grammaticale) ainsi qu'une flexion qui correspond à celle de la tête. L'analyse en elle-même est décomposée en deux étapes : une étape d'étiquetage des mots puis une étape de segmentation (figure 1).

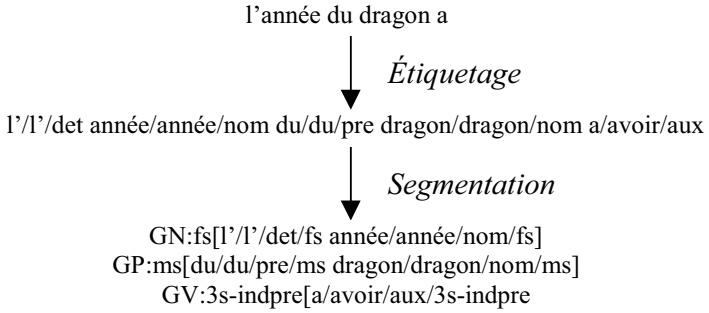

Figure 1 : Les étapes successives de l'analyseur

1) Etiquetage. Dans le cadre de la modélisation probabiliste, le processus d'étiquetage revient à aligner une séquence de mots $W = w_1,...,w_N$ et une autre de *tags* $T = t_1,...,T_N$, et à rechercher la séquence d'étiquettes T qui maximise la probabilité conditionnelle d'association. Dans SibyMot, nous avons utilisé le modèle n-POS, avec $n = 3$. SibyMot travaillant sur les lemmes, le modèle n-POS a été adapté pour estimer non plus la probabilité d'apparition d'un mot mais celle d'un lemme. Par rapport au jeu d'étiquettes de l'action GRACE (Rajman et al., 1997),

notre jeu d'étiquettes est plus réduit (une centaine d'étiquettes). En particulier, les étiquettes ne contiennent pas d'informations flexionnelles, ce qui facilite la tâche de l'étiqueteur. L'évaluation de l'analyseur a ainsi donné un taux de 97,9 % de mots correctement étiquetés sur un extrait du journal Le Monde d'environ 50 000 mots.

2) Segmentation. Pour réaliser la segmentation, nous proposons une solution originale qui s'inspire du modèle n-POS. Dans la modélisation adoptée, la phrase est vue non plus comme une séquence de mots, mais comme une séquence de *chunks* $C = c_1, \ldots, c_N$. Chaque *chunk* c_j contient un ou plusieurs mots représentés par leur étiquette grammaticale. Par analogie au modèle n-POS, la liste des parties du discours est identifiée à la liste des différentes classes de *chunk* (*GN*, *GV*, etc.) et l'ensemble des éléments d'une classe est constitué par les séquences de *tags* appartenant à cette classe (par exemple, *det_nom*, *det_nom_adj*, … pour le groupe nominal). La liste des séquences est donnée par une grammaire des *chunks,* qui contient plus de 200 000 séquences et est créée de manière automatique à partir d'une base de 200 règles sous forme d'expressions régulières. Pour la segmentation, l'évaluation a donné un taux de 93,9 % de taux de rappel sur le même corpus que précédemment.

3.2 Partie prédiction

Au sortir de l'analyseur, nous disposons d'une segmentation de la phrase en *chunks* et d'un étiquetage en classes grammaticales. Cette structure est ensuite utilisée par le module de prédiction pour établir une probabilité des mots du lexique. Sans entrer dans les détails de la réalisation (Schadle, 2003), le processus de prédiction est réalisé en cinq étapes (figure 2).

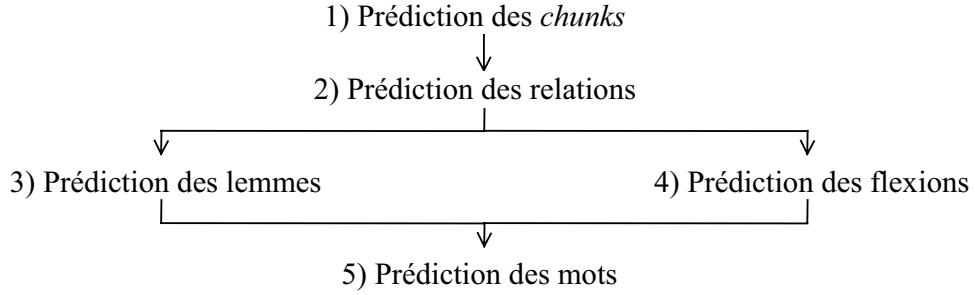

Figure 2 : Étapes de la prédiction

1) Prédiction des *chunks*. La première étape de la prédiction est chargée de fournir l'ensemble des segmentations possibles pour le mot à venir. Elle s'appuie sur la même grammaire des *chunks* que l'analyseur et donne une estimation de la probabilité de chacune d'elles. Les segmentations produites fournissent les étiquettes grammaticales du mot et du *chunk*. Par exemple, après la séquence « GN[l'année] », cette étape détermine les probabilités de « GN[l'année <adjectif cardinal> », « GN[l'année] GV[<verbe conjugué> », etc.

2) Prédiction des relations. L'objectif de cette étape est de mettre en relation le dernier *chunk* avec les *n-1 chunks* précédents. Le but implicite est de capter les relations entre syntagmes. Cette mise en relation est également probabilisée et utilise uniquement l'étiquette attribuée aux *chunks*. Sur l'exemple « GN[l'année] GP[du dragon] GV[a <participe passé> », une forte probabilité sera ainsi attribuée à la relation entre le GV et le GN.

3) Prédiction des lemmes. À ce stade, la prédiction dispose de toutes les informations nécessaires pour estimer les probabilités des lemmes. L'estimation combine une probabilité tri-lemme (à l'image du n-gramme) et une probabilité fondée sur les têtes de *chunks*. C'est cette dernière qui sur l'exemple « GN[l'année] GP[du dragon] GV[a <participe passé> » et la relation GV-GN, permet d'obtenir une forte probabilité pour « commencé ».

4, 5) Parallèlement à l'estimation des lemmes, l'étape de prédiction des flexions délivre une estimation pour chaque flexion. Grâce aux relations établies en 2) un mécanisme d'accords entre *chunks* est rendu possible. Au final, à partir des estimations des lemmes et des flexions, ces probabilités sont combinées pour calculer la probabilité des mots du lexique de SibyMot.

4 Apprentissage

Pour l'acquisition des paramètres du modèle, le corpus d'apprentissage doit être annoté. À chaque mot doit correspondre son lemme, sa catégorie grammaticale, sa flexion. Les phrases doivent être segmentées et les segments mis en relation. Nous ne disposons malheureusement pas d'un tel corpus, l'apprentissage a donc été réalisé sur un corpus non annoté manuellement. Le corpus utilisé contient près de deux millions de mots (un mois du journal Le Monde). Pour l'étiqueteur, l'apprentissage a été réalisé sur un étiquetage produit par l'analyseur Cordial. L'acquisition des paramètres des modules supérieurs a été obtenue par apprentissage non supervisé. Nous reviendrons sur cet apprentissage sous-optimal lors des résultats de l'évaluation. Quant au lexique il est extrait de ceux de l'ABU et de Lexique (accessibles sur l'internet), enrichis des données d'apprentissage et contient plus de 50 000 lemmes.

5 Évaluation

L'évaluation adoptée est proche de celle proposée dans (Bimbot et al., 1997) qui permet de comparer des modèles probabilistes à des modèles non probabilistes. Il s'agit d'une adaptation du jeu de Shannon qui consiste à proposer à partir d'un contexte, une liste de mots candidats. Chacun des mots candidats étant affecté d'un poids, la qualité du modèle est évaluée à partir de la moyenne géométrique des poids accordés à la solution correcte pour chaque contexte. Notre métrique se rapproche de cette dernière et est plus adaptée à l'évaluation des systèmes d'aide à la communication. Après chaque lettre tapée par l'utilisateur pour composer son message, le système affiche une liste d'un certain nombre de mots (ici 5). Si le mot souhaité apparaît dans la liste, les lettres non tapées sont considérées comme économisées, sinon l'utilisateur tape la lettre suivante et le système établit une nouvelle liste de propositions. On mesure ainsi le nombre de lettres économisées par rapport au nombre de lettres du message. Lors de cette évaluation nous avons comparé notre système au modèle n-gramme (ordres de 1 à 3). Le corpus d'apprentissage est le même pour les différents modèles comparés. Le corpus de test contient 50 889 mots. Les résultats obtenus sont donnés dans le tableau ci-dessous.

Modèle	1-gramme	2-gramme	3-gramme	SibyMot
% économisés	43,9 %	51,2 %	55,8 %	57,1 %

Tableau 1 : Évaluation comparée du modèle SibyMot

Les résultats montrent que notre modèle obtient des performances supérieures à l'uni-gramme (+ 13,2 %), au bi-gramme (+ 5,9 %) et au tri-gramme (+ 1,3 %). Cette dernière comparaison montre que notre modèle avec des connaissances syntaxiques obtient de meilleurs résultats qu'un modèle n-gramme simple. De plus, nous pensons que l'apprentissage a été réalisée de manière sous optimale et que le modèle dispose ainsi d'une certaine marge de progression.

6 Conclusion

Nous avons présenté dans cet article les principes du modèle de langage utilisé par le système Sibylle. Dans le cadre de la modélisation probabiliste du langage, nous proposons d'améliorer les capacités prédictives du modèle n-gramme en captant des dépendances à plus longue distance avec des *chunks*. Les résultats obtenus montrent ainsi que les capacités de notre modèle sont supérieures à celle du modèle n-gramme. Ce modèle appelé SibyMot est actuellement intégré dans l'application Sibylle, un système d'aide à la communication pour les personnes handicapées. Cette application est utilisée au CMRRF de Kerpape par des Infirmes Moteurs Cérébraux. Le module SibyMot va être également commercialisé dans un autre système d'aide à la communication par la société Microvocal. Enfin, notons que le modèle SibyMot, dans sa partie analyseur, participe à la campagne d'évaluation EASY des analyseurs syntaxiques du français, dans le cadre de l'action Technolangue.

Références

ABNEY S. (1991), Parsing by chunks. In R. Berwick, S. Abney, and C. Tenny (Eds.), *Principle based parsing*, Kluwer Academic.

BIMBOT F., EL-BÈZE M., JARDINO M. (1997), An alternative scheme for perplexity estimation. Proc. of *the International Conference on Acoustics, Speech and Signal Processing*, Munich.

BOISSIERE P. (2000) VITIPI : Un système d'aide à l'écriture basé sur un principe d'auto-apprentissage et adapté à tous les handicaps moteurs. Actes de *Handicap'00*, pp 81-86, Paris.

CERF-DANON H., EL-BÈZE M. (1991), Three different probabilistic language models: Comparison and combination. In Proceeding of *ICASSP-91*, pp 297-300, Toronto, Canada.

JELINEK F. (1976), Continuous speech recognition by statistical models. Proc. of *the IEEE*.

MAUREL D., LE PÉVÉDIC B. (2001), The syntactic prediction with Token Automata: Application to HandiAS system. *Theoretical Computer Science, vol. 267*, pp 121-129.

RAJMAN M., LECOMTE J., PAROUBEK P. (1997), Format de description lexicale pour le français. Partie 2 : Description morpho-syntaxique, *réf. GRACE GTR-3-2.1*.

SCHADLE I., LE PEVEDIC B., ANTOINE J.-Y., POIRIER F. (2001), SibyLettre : prédiction de lettre pour l'aide à la saisie de texte. Actes de *TALN'2001, vol. 2*, pp 233-242, Tours, France.

SCHADLE I. (2003), Sibylle : Système linguistique d'aide à la communication pour les personnes handicapées. *Thèse de doctorat*, Université de Bretagne Sud.