Université de Montréal
Département d'Informatique et de Recherche Opérationnelle
Intelligence Artificiel Avancé – IFT6010
Session Hiver 2003

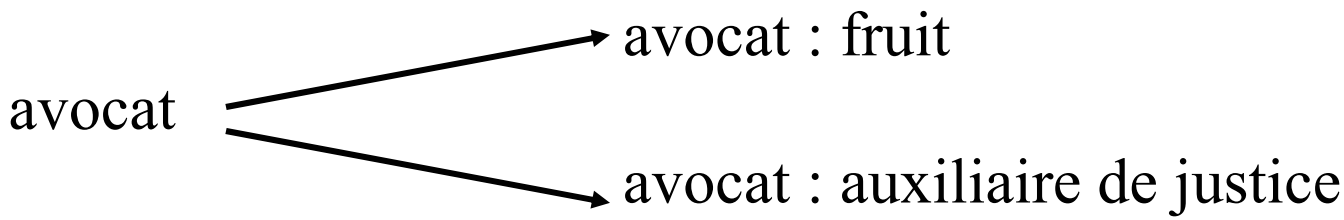
Désambiguïsation de mots

Professeur : M. Philippe Langlais

Présentée par : Leila Arras

Introduction

La désambiguïsation et le fait de pouvoir déterminer le sens d'un mot invoqué dans un contexte particulier.



Applications

- Traduction automatique :

Exemple : grille (railling, gate, bar, scale …)

- Recherche d'information :

Exemple : - court en relation avec royauté

- court en relation avec le droit

- Analyse du contenu

Exemple : L'étagère plie sous les livres.

- Analyse syntaxique

Exemple : L'homme observe la femme avec la caméra

Différentes approches

Toutes les méthodes tendent à déterminer la meilleur association entre le contexte courant et les sources d'informations trouvées(contexte, connaissances externes, encyclopédie, dictionnaires …) dans le but d'assigner un sens à chaque mots ambiguë.

Approche supervisée

Dans ce cas, un corpus non ambigu est disponible avec un ensemble d'exemples ou chaque occurrence du mot ambiguë w est annoté avec une étiquette sémantique s_k qui représente le sens contextuelle de w.

Le but est de construire un classificateur capable de classer les nouveaux cas en se basant sur le contexte d'utilisation dans le corpus utilisé.

Classification bayesienne

Tient compte des mots entourant le mot ambiguë dans une assez contexte assez large représenté par une fenêtre.

Exemple :

medication : prices, prescription, patent, increase, pharmaceutical

illegal substance : abuse, alcohol, cocaine, traffickers

Classification bayesienne

1- Apprentissage
Pour chaque sens s_k de w faire
    Pour chaque mot vj du vocabulaire faire
        $P(vj / s_k) = C(vj, s_k) / C(vj)$
    Fin
Fin
Pour chaque sens s_k de w faire
    $P(s_k) = C(s_k) / C(w)$
2- Désambiguïsation
Pour chauqe sens s_k de w faire
    $score(s_k) = \log P(s_k)$
    Pour chauqe mot vj dans la fenêtre de contexte C faire
        $score(s_k) = score(s_k) + \log P(vj / s_k)$
    Fin
Fin
Choisir s' = argmax s_k $score(s_k)$

Information.-théorique

Trouver une caractéristique textuel qui permet d'indiquer de manière fiable le sens utilisé d'un mot ambigu donné.

Exemple :

Mot ambiguë	**Indicateur**	**Valeur** → **Sens**
Prendre	objet	mesure → to take
		décision → to make
vouloir	temps	présent → to want
		conditionnel → to like

Information.-théorique

Trouvez partition aléatoire P = {P1,P2} de $\{t_1,\ldots,t_m\}$

Tant que(amélioration) faire
    trouver partition Q = {Q1,Q2} de $\{x_1,\ldots,x_n\}$
        qui maximise I{P;Q}
    trouver partition P = {P1,P2} de $\{t_1,\ldots,t_m\}$
        qui maximise I{P;Q}
Fin

Traduire Prendre avec :
$\{t_1,\ldots,t_m\}$={take,make,rise,speak};
$\{x_1,\ldots,x_n\}$={mesure,note,exemple,décision,parole}
P1={take,rise}; P2={make,speak}
Q1={mesure,make,exemple}; Q2={décision,parole}

Utilisation .de dictionnaire

Aucune information disponible sur la catégorisation d'une instance spécifique d'un mot. D'où le besoin de définitions de mot dans les dictionnaires.

Lesk(1986) est parti de l'idée simple que les définitions d'un mot du dictionnaire sont considérées comme étant les indicateurs les plus probables du sens d'un mot.

Etant donné un contexte :C
Pour tous les sens s_k de w faire
    Score(s_k) = chevauchement(D_k,$U_{vj\ in\ C}$ Evj)
Fin
Choisir s'=argmax s_k score(s_k)

Utilisation .de thesaurus

Exploite la catégorisation sémantique faite par le thesaurus.

Les catégories sémantiques des mots dans un contexte déterminent la catégorie sémantique du contexte en entier ce qui nous détermine le sens du mot ambiguë

Textes trad.uits

Cette approche consiste à utiliser un corpus source comme étant la premìre langue, contant le ou les mots ambiguës, un autre corpus parallèle dans une deuxième langue.

Exemple :

	Sens1	**Sens2**
Definition	legal share	attention, concern
Traduction	beteiligung	intenresse
Collocation anglaise	acquérir un intérêt	montrer un intérêt
Traduction	beteilingung erwerben	interesse zeigen

Un sens par discours

Le sens d'un mot est extrêmement cohérent avec le document ou il se trouve.

Un sens par cooccurrence

Les mots entourant un mot ambiguë fournissent de bonne indications sur le sens de ce mot(distance relative, trigramme…).

Approche non supervisée

On ne possède aucune information sur la classification des données au niveau des exemples d'apprentissage.

Le but est de regrouper les contextes d'un mot ambiguë en un certain nombre de groupes et ainsi différencier entre ces groupes sans les avoir tagger au préalable.

SENSEVAL

Un système pouvant évaluer des applications qui détermine le sens des mots de manière automatique.

Ce système a pris place dans ACL SIGLEX(the Lexicons Special Interest Group of the Association for Computational Linguistics) en 1998.

Conclusion

Avec des moyens d'évaluation minutieux, il serai possible de cerner et de comprendre les forces et faiblesses des algorithmes de désambiguïsation de mots et penser ainsi à améliorer ceux qui existent et déduire de nouvelles approches.

Bibliographie

1-Nancy Ide and Jean Véronis Computational linguistics, 1998 24(1)

2-David Yarowsky, Unsupervised word sense disambiguation rivaling supervised methods

3-Christopher D.Manning and Hinrich shutze , Foundations of statistical natural language processing

4-Adam Kilgarriff, SENSEVAL : An exericise in Evaluating word sense disambiguation programs, may1998