L'extraction de lexiques bilingues

Carmen Alvarez

IFT6010

31 Mars 2003

Les types de corpus multilingues

1. Corpus parallèles : *Hansard*
   - Correspondance de l'ordre des phrases et mots
   - Corrélation entre les fréquences des mots
2. Corpus comparables : *The Aarhus Corpus of Contract Law* (français, anglais, danois)
   - Origine, domains, thèmes similaires
   - Vocabulaire originel
3. Corpus dans différentes langues : *European Corpus Initiative*, collection de journaux français, anglais, allemands, espagnols, et danois

Amélioration de la sélection de traduction, Zhou, Ding, Huang, 2001.

Traduction de chinois vers anglais.

Ressources :

- Corpus monolingues indépendants
- Lexique chinois-anglais de base

Processus :

- Traiter chaque corpus indépendamment et construire une base de données de dépendances pour chaque langue
- Construire un lexique chinois-anglais à partir du lexique de base et les bases de données de dépendances

Resultat : Utiliser le lexique étendu avec un modèle de langue pour réaliser la traduction

Modèle de dépendances

Triple de dépendance :

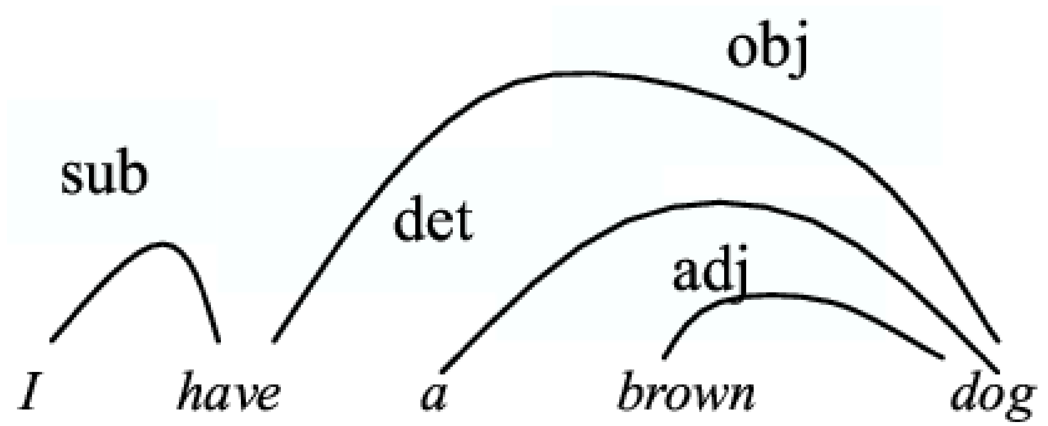
$$trp = (m1, rel, m2) \tag{1}$$

où $m1$ est le mot tête, $m2$ est le mot dépendant, et rel est la relation de dépendance.

La relation de dépendance peut être :

- *sujet-verbe*
- *verbe-objet*
- *adjectif-verbe*
- etc.

Modèle de dépendances : Un exemple



Triples de dépendance : *(have, subj, I), (I, sub-of, have), (have, obj, dog), (dog, obj-of, have), (dog, adj, brown), (brown, adj-of, dog), (dog, det, a), (dog, det-of, a)*

Similarité des dépendances

Hypothèse : Les relations des triples de dépendance du chinois et de l'anglais ont une forte correspondance. (La relation *subj-verb* dans un triple chinois est souvent traduite á une relation *subj-verb* en anglais.)

Exemple :

- $trp_C = (A_c, D_c, B_c)$
- $trp_E = (A_e, D_e, B_e)$
- A_e et B_e sont les traductions des mots A_c et B_c, et on peut facilement déterminer D_e à partir de D_c.

Les auteurs ont verifié l'hypothèse avec un petit corpus parallèle.

Similarité entre mots : Cas monolingue

L'idée : La similarité entre deux mots peut être calculée à partir du contexte de chaque mot.

Représentation du contexte : Un mot m est répresenté par un vecteur de co-occurrence $\vec{v} =< v_0, v_1, ..., v_L >$, de taille L, où L est la taille du lexique et v_i est la fréquence de co-occurrence dans le corpus entre m et le mot m_i du lexique.

- La similarité entre deux mots m_1 et m_2 est déterminé par le calcul de similarité entre les vecteurs de co-occurrence $\vec{v_1}$ et $\vec{v_2}$.

Similarité entre mots : Cas monolingue

		étudiant	grève	professeur	travail
$\vec{v}_{ecole}$	$=<$	543,	0,	223,	23$>$
$\vec{v}_{universite}$	$=<$	498,	53,	260,	109$>$
$\vec{v}_{syndicat}$	$=<$	0,	676,	12,	334$>$

On peut combiner cette approche avec le modèle de dépendance en utilisant un vecteur $\vec{v}$ où chaque élement v_i est la fréquence ou la probabilité du triple de dépendance (m, rel, m_i). On aura besoin d'un tel vecteur pour chaque type de relation.

Similarité entre mots : Cas monolingue

Le triple $(m1, rel, m2)$ représente 3 évènements :

1. un mot aléatoire $m1$
2. une dépendance aléatoire rel
3. un mot aléatoire $m2$

Similarité entre mots : Cas monolingue

L'information du triple est donnée par :

$$
\begin{aligned}
I(m1, rel, m2) &= \log(P(A,B,C)) - \log(P(B)P(A|B)P(C|B)) \\
&= \log \frac{P(A,B,C)}{P(B)P(A|B)P(C|B)} \\
&= \log \frac{|m1, rel, m2| \times |*, rel, *|}{|m1, rel, *| \times |*, rel, m2|}
\end{aligned}
$$

où :

- $P(A|B) = \frac{|(m1,rel,*)|}{|(*,rel,*)|}$;$P(C|B) = \frac{|(*,rel,m2)|}{|(*,rel,*)|}$
- $P(B) = \frac{|(*,rel,*)|}{|(*,*,*)|}$
- $P(A,B,C) = \frac{|(m1,rel,m2)|}{|(*,*,*)|}$

Similarité entre mots : Cas monolingue

Soit

$$T(m) = \{< rel, m' > | I(m, rel, m') > 0\} \tag{2}$$

les relations rel et mots m' où il y a un triple avec le mot m.

Alors, la similarité entre les mots $m1$ et $m2$ est donnée par

$$Sim(m1, m2) = \frac{\sum_{(rel,m) \in T(m1) \cap T(m2)} (I(m1, rel, m) + I(m2, rel, m))}{\sum_{(rel,m) \in T(m1)} I(m1, rel, m) + \sum_{(rel,m) \in T(m2)} I(m2, rel, m)} \tag{3}$$

Similarité entre mots : Cas bilingue

$$T(m_C) = \{< rel_C, m'_C > | I(m_C, rel_C, m'_C) > 0\}$$

$$T(m_E) = \{< rel_E, m'_E > | I(m_E, rel_E, m'_E) > 0\}$$

Alors, la similarité entre les mots m_C et m_E est donnée par :

$$Sim(m_C, m_E) = \frac{I_{common}(m_C, m_E)}{\sum_{(rel_C, m'_C) \in T(m_C)} I(m_C, rel_C, m'_C) + \sum_{(rel_E, m'_E) \in T(m_E)} I(m_E, rel_E, m'_E)} \quad (4)$$

où I_{common} est calculée par 3 méthodes :

1. Chinois vers anglais
2. Anglais vers chinois
3. Chinois vers anglais et anglais vers chinois

Similarité entre mots : Cas bilingue

Méthode 1 : Chinois vers Anglais.

$$
\begin{aligned}
T_{C\to E}(m_E|m_C) = \{(rel_E, m'_E) \quad | \quad & (rel_C, m'_C) \in T(m_C), \\
& rel_E = f_{reln}(rel_C), \\
& m'_E \in Tran(m'_C)\} \\
& \cap T(m_E)
\end{aligned}
$$

$$
\begin{aligned}
T_{C\to E}(m_C|m_E) = \{(rel_C, m'_C) \quad | \quad & (rel_E, m'_E) \in T(m_E), \\
& rel_E = f_{reln}(rel_C), \\
& m'_E \in Tran(m'_C)\}
\end{aligned}
$$

Similarité entre mots : Cas bilingue

Méthode 1 : Chinois vers Anglais.
Alors,
$Sim_{C\rightarrow E}(m_C, m_E) =$

$$\frac{\sum_{(rel_C, m'_C)\in T_{C\rightarrow E}(m_C|m_E)} I(m_C, rel_C, m'_C) + \sum_{(rel_E, m_E)\in T_{C\rightarrow E}(m_E|m_C)} I(m_E, rel_E, m'_E)}{\sum_{(rel_C, m'_C)\in T(m_C)} I(m_C, rel_C, m'_C) + \sum_{(rel_E, m'_E)\in T(m_E)} I(m_E, rel_E, m'_E)} \quad (5)$$

La sélection de traduction

Le but est de traduire le triple chinois $trp_C = (m_{C1}, rel_C, m_{C2})$ vers un triple anglais $trp_E = (m_{E1}, rel_E, m_{E2})$.

$$trp_E =^{\text{argmax}}_{trp_E} P(trp_E | trp_C)$$

On applique le théorème de Bayes :
$P(trp_E | trp_C) = \frac{P(trp_E)P(trp_C|trp_E)}{P(trp_C)}$
Donc,

$$trp_E =^{\text{argmax}}_{trp_E} P(trp_E)P(trp_C|trp_E)$$

- $P(trp_E)$ est determinée par le modèle de langue
- $P(trp_C|trp_E)$ est determinée par le modéle de dépendance :

$$P(trp_C|trp_E) = P(m_{C1}|rel_C, trp_E) \times P(m_{C2}|rel_C, trp_E) \times P(rel_C|trp_E)$$

où $trp_C = (m_{C1}, rel_C, m_{C2})$.

La sélection de traduction

On définit la vraisemblance d'un triple :

$$\mathfrak{L}(trp_C|trp_E) =$$
$$Sim_{E\to C}(m_{C1}, m_{E1}) \times Sim_{E\to C}(m_{C2}, m_{E2}) \times P(rel_C|rel_E)$$

C'est comme une probabilité, mais pas bornée entre [0,1].

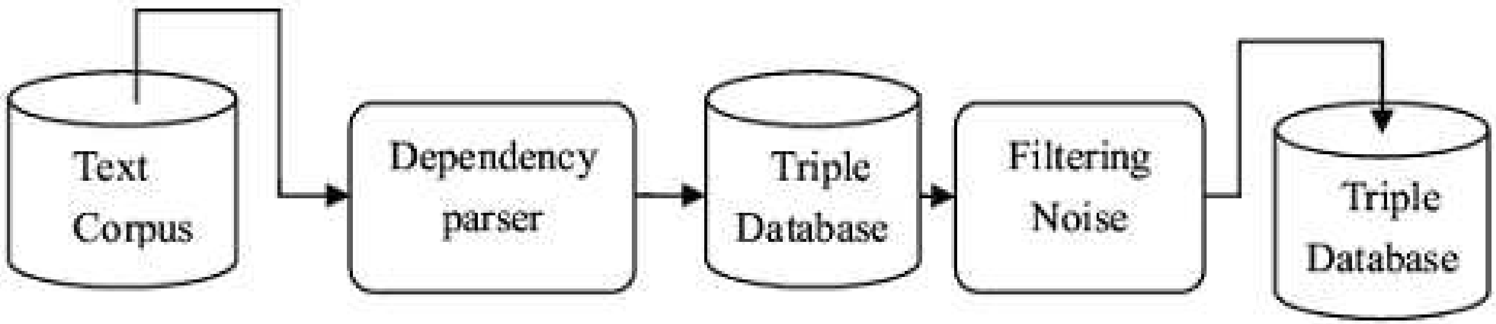
Mais $P(rel_C|trp_E) = P(rel_C|rel_E) \approx 1$. Alors,

$$\mathfrak{L}(trp_C|trp_E) = Sim_{E\to C}(m_{C1}, m_{E1}) \times Sim_{E\to C}(m_{C2}, m_{E2})$$

Donc,

$$\begin{aligned} trp_E &= \underset{trp_E}{\operatorname{argmax}}\, P(trp_E)P(trp_C|trp_E) \\ &= \underset{trp_E}{\operatorname{argmax}}\, P(trp_E) \times \mathfrak{L}(trp_C|trp_E) \\ &= \underset{trp_E}{\operatorname{argmax}}\, P(trp_E) \times Sim_{E\to C}(m_{C1}, m_{E1}) \times Sim_{E\to C}(m_{C2}, m_{E2}) \end{aligned}$$

L'entrainment du modèle



1. Traiter chaque corpus indépendamment
2. Chaque fois un triple est vu, l'ajouter à la base de donées de dépendences où incrémenter sa fréquence
3. Filtrer les triples "bruités" (pas réalisé par les auteurs).

Après 50 heures d'entrainment pour les deux langues :

- 19 millions de triples anglais
- 33 millions de triples chinois

L'entrainment du modèle

- Lexique anglais <-> chinois de base : 75000 entrées.
- 3 dépendances ont été testées :
  - Verb-object
  - Verb-adverb
  - Subject-verb

La matrice de similarité :

	E_1	E_2	...	E_m
C_1	S_{11}	S_{12}	...	S_{1m}
C_2	S_{21}	S_{22}	...	S_{2m}
...	...	...	...	...
C_n	S_{n1}	S_{n2}	...	S_{nm}

Tests et évaluation

3 modèles et 3 tests, où le but est de trouver le meilleur triple anglais $trp_E = (m_{E1}, rel_E, m_{E2})$, étant donné le triple chinois $trp_C = (m_{C1}, rel_C, m_{C2})$:

1. Modèle A. Modèle de traduction seulement.

$$trp_E = \left(\operatorname*{argmax}_{m_{E1} \in Trans(m_{C1})} |m_{E1}|, verb{-}object, \operatorname*{argmax}_{m_{E2} \in Trans(m_{C2})} |m_{E2}| \right)$$

2. Modèle B. Modèle de dépendance seulement.

$$trp_E = \operatorname*{argmax}_{trp_E} P(trp_E) = \operatorname*{argmax}_{\substack{m_{E1} \in Trans(m_{C1}) \\ m_{E2} \in Trans(m_{C2})}} P(m_{E1}, verb-obj, m_{E2})$$

3. Modèle C. Modèle de traduction et dépendance.

$$trp_E = \operatorname*{argmax}_{trp_E} P(trp_E) = \operatorname*{argmax}_{\substack{m_{E1} \in Trans(m_{C1}) \\ m_{E2} \in Trans(m_{C2})}} P(m_{E1}, verb-obj, m_{E2}) \times sim_{E \to C}(m_{C1}, m_{E1}) \times sim_{E \to C}(m_{C2}, m_{E2})$$

Tests et évaluation

1. Test 1. Verbes de haute fréquence

Model	#Correct	Percentage
Model A	393	53.8%
Model B	512	70.1%
Model C	519	71.1%

2. Test 2. Verbes de basse fréquence

Model	#Correct	Percentage
Model A	61	56.5%
Model B	85	78.7%
Model C	88	81.5%

3. Test 3. Triples de basse fréquence

Model	#Correct	Percentage
Model A	182	53.5%
Model B	283	83.2%
Model C	289	85.0%

Une amélioration

Le lexique de base est parfois limité. Un exemple : Le mot chinois 展开 est traduit par le lexique de base comme "unfold" et "develop", mais le sens change quand le mot est utilisé dans le triple (展开, *verb-object*, 进攻), qui est traduit par "*launch, verb-object, attack.*"

Une amélioration possible est de faire une consultation cyclique dans le lexique de bas, pour etendre les mots à considerer pour la traduction. La fonction de traduction devient :

$Tran1(x) =$

$$\{x'''|x''' \in Tran(x''), x'' \in Tran(x'), x' \in Tran(x)\} \cup Tran(x)$$

Ce modèle a donné une amélioration de 1% sur la performance du Modèle C.

Références

[1] Abaitua, J. 2000. Tratamiento de corpora bilingües. *Fundación Duques de Soria, Seminario "La ingeniería lingüística en la sociedad de la información."*. Universidad de Duesto, Espagne.

[2] Fung, P. 1998. A statistical view on bilingual lexicon extraction from parallel to non-parallel corpora. *Third Conference of the Association for Machine Translation in the Americas*. Springer. 1-16.

[3] Lin, D. 1998. Automatic retrieval and clustering of similar words. *COLING ACL98* Montréal, Canada.

[4] Zhou, M., Ding, Y., Huang, C. 2001. Improving translation selection with a new translation model trained by independent monolingual corpora. *Computational linguistics and Chinese Language Processing. Vol. 6, No.1* 1-26.