PARADOCS
Un système d'identification automatique de documents parallèles

Alexandre Patry et Philippe Langlais

RALI
Département d'informatique et de recherche opérationnelle
Université de Montréal

TALN — 6-10 Juin 2005

Plan

Documents parallèles ?

Définition possible

Deux documents sont parallèles s'ils véhiculent le même contenu dans le même ordre.

Alignés au niveau des phrases (bitexte), ils servent à créer des ressources ou à déployer des applications telles que :

- lexiques, fiches terminologiques
- traduction
- recherche d'information (trans-linguistique)
- détection de paraphrases
- concordanciers bilingues (ex : TSRALI.COM)

Lire (Véronis,2000)

Corpus parallèles

- Quelques corpus parallèles existent :
  - Débats parlementaires canadiens (français, anglais, inuktitut)
  - Débats parlementaires de Hong-Kong (anglais, chinois)
  - Débats parlementaires européens (français, italien, espagnol, portugais, anglais, allemand, hollandais, danois, suédois, grecque, finnois)
  - Bible, Coran, Harry Potter, etc.
- Mais ils sont cependant peu nombreux et souvent peu adaptés
- Pourquoi ne pas aller en chercher sur internet ?

Problème auquel nous nous intéressons

Soient un ensemble de documents dans une langue source et un ensemble dans une langue cible, nous voulons trouver les paires de documents parallèles.

Scénario typique

1. Télécharger un site web bilingue.
2. Identifier la langue de chaque document à l'aide d'un outil comme `SILC`.
3. Détecter les paires de documents parallèles.
4. Aligner les phrases des documents parallèles.
5. Entraîner un engin de traduction statistique.

Indices structurels

- Utiliser les noms de fichiers où de liens

Exemple

http ://www.gc.ca/main_f.html
http ://www.gc.ca/main_e.html

http ://www2.ville.montreal.qc.ca/plan-urbanisme/index.shtm
http ://www2.ville.montreal.qc.ca/plan-urbanisme/en/index.shtm

http ://applicatif.ville.montreal.qc.ca/fr/commfr.asp ?id=2993
http ://applicatif.ville.montreal.qc.ca/en/comman.asp ?id=2994

- Problèmes : politique des noms de fichiers non standardisée, traductions incomplètes, mauvaises, ou non maintenues
- Il semble donc qu'il faille inspecter le contenu des documents

Approche avec lexique bilingue

- **Idée :** apparier ensemble les documents qui partagent le plus de mots selon un lexique.
- Pour chaque document source s, nous voulons le document cible t maximisant :

$$\frac{\text{Nombre de mots que } s \text{ et } t \text{ partagent selon le lexique}}{\text{Nombre de mots dans } s + \text{Nombre de mots dans } t}$$

Exemple

Resumption of the session

Reanudación del período de sesiones $\left(\frac{3}{9}\right)$

Declaro reanudado el período de sesiones del Parlamento Europeo , interrumpido el viernes 17 de diciembre pasado. $\left(\frac{5}{20}\right)$

Approche avec lexique bilingue

- **Idée :** apparier ensemble les documents qui partagent le plus de mots selon un lexique.
- Pour chaque document source s, nous voulons le document cible t maximisant :

$$\frac{\text{Nombre de mots que } s \text{ et } t \text{ partagent selon le lexique}}{\text{Nombre de mots dans } s + \text{Nombre de mots dans } t}$$

Exemple

Resumption of the session

Reanudación del período de sesiones $\left(\frac{3}{9}\right)$

Declaro reanudado el período de sesiones del Parlamento Europeo , interrumpido el viernes 17 de diciembre pasado. $\left(\frac{5}{20}\right)$

Classification d'une paire

Un classificateur de type AdaBoost (Freund and Schapire, 1999) est entraîné et utilisé pour identifier une paire (de documents) comme parallèle ou non.

- **Entrée** : un vecteur caractérisant une paire de documents
- **Sortie** : $\oplus$ ou $\ominus$
- Nous utilisons un réseau de neurone à une couche cachée comme classificateur faible
- 75 itérations

Le lexique est-il nécessaire ?

Quiz

The Legislative Assembly convened at 3.30 pm.	sitamiq, ipuru 1, 1999
Mr. Quirke (Clerk-Designate) :	maligaliurvik matuiqtaulauqtuq 3 :30mi unnusakkut
THURSDAY, APRIL 1, 1999	mista kuak (titiraqti - tikkuaqtausimajuq) :

Le lexique est-il nécessaire ?

Quiz

The Legislative Assembly convened at 3.30 pm.	sitamiq, ipuru 1, 1999
Mr. Quirke (Clerk-Designate) :	maligaliurvik matuiqtaulauqtuq 3 :30mi unnusakkut
THURSDAY, APRIL 1, 1999	mista kuak (titiraqti - tikkuaqtausimajuq) :

Caractéristiques lexicales observées

Nous considérons trois ensembles de caractéristiques lexicales dans cette étude :

- **Nombres**
  Toute séquence de un ou plusieurs chiffres.
- **Ponctuations**
  Parenthèses, crochets et guillemets.
- **Entités nommées**
  Mot ne débutant pas une phrase et dont la première lettre est en majuscule.

Représentation vectorielle d'un document

Exemple

Approximately 60% very roughly, 60% to 40%, when the 60% is paid by the tenant and 40% is approximately paid by the Government subsidy.

apiqqutiqaqqaujunga akunialuk, angiqqaugaluarakku $60 milian kaivainnaqtuq kiinaujaqarvingmut, kisianittauq tusaqtitauvalliaqqaugama, takuvallialiqtugu $39 milian 807 tausan ammalu taanna angiqtauguni taikkuali amiakkujut $60 milianut tikillugu kisumut atuqtaugajaqpat ?

Les vecteurs sont donc $(0_{39}, 2_{40}, 3_{60}, 0_{807})$ et $(1_{39}, 0_{40}, 2_{60}, 1_{807})$.

Mesure de cosinus

- Document = vecteur
- Similarité entre deux documents :

$$\cos(v_1, v_2) = \frac{v_1 \cdot v_2}{||v_1|| \, ||v_2||}$$

Exemple

La similarité de $(0, 2, 3, 0)$ et $(1, 0, 2, 1)$ est :

$$\cos(v_1, v_2) = \frac{0 \cdot 1 + 2 \cdot 0 + 3 \cdot 2 + 0 \cdot 1}{3.6 \cdot 2.4}$$

$$= \frac{6}{8.8} = 0.68$$

Représentation séquentielle d'un document

Exemple

Approximately 60% very roughly, 60% to 40%, when the 60% is paid by the tenant and 40% is approximately paid by the Government subsidy.

apiqqutiqaqqaujunga akunialuk, angiqqaugaluarakku $60 milian kaivainnaqtuq kiinaujaqarvingmut, kisianittauq tusaqtitauvalliaqqaugama, takuvallialiqtugu $39 milian 807 tausan ammalu taanna angiqtauguni taikkuali amiakkujut $60 milianut tikillugu kisumut atuqtaugajaqpat ?

Les séquences sont $< 60, 60, 40, 60, 40 >$ et $< 60, 39, 807, 60 >$.

Calcul de la distance d'édition normalisée

On compte le nombre minimal d'opérations pour transformer la première séquence en la deuxième que l'on normalise (par la longueur de la plus longue des deux séquences).

Exemple

Pour transformer $< 60, 60, 40, 60, 40 >$ en $< 60, 39, 807, 60 >$, il faut :

1. Remplacer le deuxième 60 par 39 $< 60, 39, 40, 60, 40 >$
2. Remplacer le premier 40 par 807 $< 60, 39, 807, 60, 40 >$
3. Supprimer le dernier 40 $< 60, 39, 807, 60 >$

Distance d'édition de 3 normalisée par la longueur de la plus longue séquence (5) donne une distance d'édition normalisée de 0.6.

Calcul de la distance d'édition normalisée

On compte le nombre minimal d'opérations pour transformer la première séquence en la deuxième que l'on normalise (par la longueur de la plus longue des deux séquences).

Exemple

Pour transformer $< 60, 60, 40, 60, 40 >$ en $< 60, 39, 807, 60 >$, il faut :

1. Remplacer le deuxième 60 par 39 $< 60, 39, 40, 60, 40 >$
2. Remplacer le premier 40 par 807 $< 60, 39, 807, 60, 40 >$
3. Supprimer le dernier 40 $< 60, 39, 807, 60 >$

Distance d'édition de 3 normalisée par la longueur de la plus longue séquence (5) donne une distance d'édition normalisée de 0.6.

Calcul de la distance d'édition normalisée

On compte le nombre minimal d'opérations pour transformer la première séquence en la deuxième que l'on normalise (par la longueur de la plus longue des deux séquences).

Exemple

Pour transformer $< 60, 60, 40, 60, 40 >$ en $< 60, 39, 807, 60 >$, il faut :

1. Remplacer le deuxième 60 par 39 $< 60, 39, 40, 60, 40 >$
2. Remplacer le premier 40 par 807 $< 60, 39, 807, 60, 40 >$
3. Supprimer le dernier 40 $< 60, 39, 807, 60 >$

Distance d'édition de 3 normalisée par la longueur de la plus longue séquence (5) donne une distance d'édition normalisée de 0.6.

Calcul de la distance d'édition normalisée

On compte le nombre minimal d'opérations pour transformer la première séquence en la deuxième que l'on normalise (par la longueur de la plus longue des deux séquences).

Exemple

Pour transformer $< 60, 60, 40, 60, 40 >$ en $< 60, 39, 807, 60 >$, il faut :

1. Remplacer le deuxième 60 par 39 $< 60, 39, 40, 60, 40 >$
2. Remplacer le premier 40 par 807 $< 60, 39, 807, 60, 40 >$
3. Supprimer le dernier 40 $< 60, 39, 807, 60 >$

Distance d'édition de 3 normalisée par la longueur de la plus longue séquence (5) donne une distance d'édition normalisée de 0.6.

Pointages d'alignement

L'aligneur JAPA apparie les phrases d'une paire de documents parallèles qui sont en relation de traduction

Exemple

Sortie de JAPA :

parallèle	non parallèle
1-1 1.02964	1-1 1.32479
1-1 1.02964	1-0 3.57413
1-1 0.380811	1-0 10.773
...	...
1-1 -126.814	0-2 5855.3
1-1 -127.277	1-2 5858.6

On retient le score global d'alignement ainsi que le ratio d'alignements `1-1`, d'insertion ou de suppression (`0-1` ou `1-0`), etc.

Expérience contrôlée

Corpus EUROPARL

- EUROPARL : débats parlementaires européens tenus entre avril 1996 et septembre 2003 (11 langues)
- Partie anglaise-espagnole : 487 textes dans chaque langue $\Rightarrow$ 237 169 paires potentielles.

Protocole d'évaluation

$$\text{précision} = \frac{\text{Nb de paires parallèles identifiées}}{\text{Nb de paires identifiées}}$$

$$\text{rappel} = \frac{\text{Nb de paires parallèles identifiées}}{\text{Nb de paires parallèles}}$$

$$\text{f-mesure} = \frac{2 \times \text{précision} \times \text{rappel}}{\text{précision} + \text{rappel}}$$

Résultats sur EUROPARL

Configuration							Performances		
COS.	EDIT	NOMBRE	PONCT	NOMS	POINT	N-M	Pré.	Rap.	f_1
	√	√	√	√			100	100	100
√	√	√	√	√	√	√	99.8	99.8	99.8
	√	√					98.3	99.8	99.0
					√		85.8	99.8	92.1
						√	65.6	99.4	77.1
					√	√	49.3	99.4	62.7
√		√	√	√			24.6	99.2	38.7
√		√					12.4	98.9	21.8

- La distance d'édition est très fiable.
- Les pointages issus de l'alignements ne sont pas de bons indicateurs.
- L'approche lexique donne également des résultats parfaits.

Expérience sur le terrain

Le corpus PAHO

Le corpus PAHO a été extrait du site de la *Pan American Health Organisation*[1] en 2002. Selon SILC, il contient 2523 fichiers anglais et 4355 fichiers espagnols, totalisant plus de 10 millions de paires potentiellement parallèles.

- documents bilingues, traductions incomplètes, etc.
- beaucoup de documents identiques ou quasi-identiques
- documents plutôt courts
- absence de référence . . .

Protocole d'évaluation

les performances d'engins de traduction entraînés sur les paires identifiées parallèles ont été utilisées comme points de comparaison.

[1] http://www.paho.org

bitext	*N*	NIST	BLEU	precision
COSINE ∪ EDIT	494	5.3125	0.2435	99.0
LEXIQUE	529	5.1989	0.2304	89.2
EDIT	390	5.1342	0.2290	99.0
COSINE	333	5.1629	0.2256	99.7

- La mesure de cosinus semble être plus efficace pour les textes courts.
- La distance d'édition et la mesure de cosinus se complètent (seulement 229 paires en commun).
- L'approche sans lexique est "meilleure"

Récapitulation

Morale

Il est possible d'identifier les documents parallèles d'un corpus sans utiliser de lexique bilingue.

- Idée initialement proposée par (Nadeau et Foster, 2004) pour apparier des dépêches.
- Nous avons vérifié l'importance de tenir compte de l'ordre des caractéristiques.
- Nous avons testé l'impact de la méthode sur une tâche réelle, soit la traduction statistique.
- Nous avons intégré un algorithme d'apprentissage (AdaBoost) à la chaîne de traitements.

Travaux reliés

- (Nadeau et Foster, 2004) ont suggéré la mesure de cosinus sur différents groupes d'unités lexicales pour identifier les documents parallèles du site `http://www.newswire.ca`.
- Complémentaire à l'approche de (Munteanu et al., 2004). Les auteurs montrent qu'il est possible d'améliorer un système de traduction en détectant des phrases parallèles dans des corpus comparables.

Travaux futurs

- Explorer de nouvelles caractéristiques (ex : cognates)
- Tester l'impact de l'intégration d'un lexique bilingue.
- Ne pas considérer systématiquement toutes les paires possibles.

Références


J. Chen and J.Y. Nie 2000. Parallel Web text mining for cross-language IR *RIAO*, Paris, France pages 62-77


Y. Freund and R.E. Schapire 1999. A short introduction to boosting In *Journal of Japanese Society for Artificial Intelligence*, Vol. 14, N. 15, pages 771-780.


Dragos Stefan Munteanu and Alexander Fraser and Daniel Marcu 2004. Improved Machine Translation Performance via Parallel Sentence Extraction from Comparable Corpora. *HLT-NAACL* pages 265-272


David Nadeau and George Foster 2004. Real-Time Identification of Parallel Texts from Bilingual News feed *Computational Linguistics in the North East*, Montréal, 2004 pages 21-36


Édt. Jean Véronis, 2000 *Parallel Text Processing, Alignment and Use of Translation Corpora*. Kluwer Academic

Réalisation typique : PTMiner

PTMiner (Chen and Nie, 2000), le système d'extraction de textes parallèles à partir du web développé au RALI procède de la manière suivante :

1. Déterminer des paires candidates en utilisant les noms des fichiers.
2. Filtrer les paires retournées en utilisant :
   - la taille des fichiers
   - la langue des fichiers
   - un lexique bilingue
   - la sortie d'un aligneur de textes au niveau des phrases