Adaptation d'un Système de Traduction Automatique Statistique

Fethi BOUGARES

Maître de conférences
Laboratoire d'Informatique de l'Université du Maine

2015 - 2016

Sommaire

Traduction automatique statistique

Traduire S (source) en T (target)

Traduire $\Rightarrow$ Trouver la meilleure traduction parmi l'ensemble des hypothèses

$$\hat{t} = \underset{t \in T}{\text{argmax}}\, P(t|s) = \underset{t \in T}{\text{argmax}}\, P(s|t)P(t)$$

$P(s|t)$: Modèle de traduction
$P(t)$: Modèle de langage
$argmax$: algorithme de recherche

Motivation : un exemple

Source : elles représentent les étendues de l'imagination humaine qui remontent à l'aube du temps.

Baseline : they represent the bodies of the human imagination back at the dawn of time.

Baseline + DA-Trans. Model : they represent the bodies of the human imagination that date back to the dawn of time.

Baseline + DA-Trans. Model + topic : they represent the bodies of the human imagination that go back the dawn of time.

Référence : *they represent branches of the human imagination that go back to the dawn of time.*

Motivation : un autre exemple

Source : le débit est en augmentation très rapide

Baseline : the speed is growing very rapidly.

Baseline + DA-Trans. Model : the throughput is rising very fast.

Baseline + DA-Trans. Model + topic : the flow is growing very rapidly

Référence : *these flows are increasing very rapidly.*

Motivation : Traduction et domaine

La plupart des données parallèles sont fournies par des organisations internationales (UN, Parlement européen).

→ Non adaptées pour construire un système de traduction pour d'autres domaines :

- Legal domain
- Information Tech
- Histoire ...

⇒ Objectif : Augmenter les connaissances sur le texte à traduire (donc améliorer la qualité de la traduction)

⇒ Méthodes : Adaptation de systèmes

C'est quoi un domaine ?

Comment définir *"le domaine"*

- des documents qui traitent le même sujet
- restrictions lexicales, syntaxiques et sémantiques
- utilisation fréquentes des même constructions
- la structure du text
- utilisation des symboles spécials

Détection automatique du domaine :

- Classification
- Regrouppement de documents de même domaine
- Critère : TF-IDF, distance entre documents

Sommaire

SMT : Données et Apprentissage

Apprentissage des modèles :

- Extraction des règles ou des séquences de mots
  $\rightarrow$ À partir des bitextes
- Apprentissage des LM
  $\rightarrow$ À partir des données monolingues

Données d'apprentissage :

- Disponible en fonction du paire de langue
- Anglais/Français - Français/Portugais
- Disponible en fonction du domaine
- News / finance / médical / scientifique / films

Un système par domaine ?

Traduction avec plusieurs systèmes (un par domaine)

- Collecter des données pour chaque domaine
- Créer un système par domaine
- Détecter le domaine de la phrase à traduire
- Traduire avec le bon système

Problème :

Il n'y a pas assez de données pour tous les domaines !
C'est ne pas pratique comme méthode

Solution : Techniques d'adaptation

- Sélection de données ;
- Adaptation de la table de traduction (pondération) ;
- Adaptation de modèle de langage ;
- Génération automatique des données

Sommaire

Sélection de données

Objectif : *"extracting sentences from a large general-domain parallel corpus that are most relevant to the target domain."* Amittai Axelrod et al.

$\Rightarrow$ Il faudra donc :

- Données *in-domain*
- Données *out-of-domain* (en grande quantité)
- Une méthode de sélection (une mesure de similarité)

Sélection de données

Utilisation des techniques de recherche d'Information :

- La requête : Données *in-domain*
- La base de données : *out-of-domain*

Lu et al. (2007) :

- Avec un système de recherche d'Information
- Calculer un score pour chaque pair de phrase
- Fonction : similarité avec les phrases de test (source)

Sélection de données

Ittycheriah and Roukos (2007) :

- *"sub-sampling"* de corpus d'apprentissage
- recouvrement $n-gram$ maximal (avec le test)
- utilisation pour l'apprentissage de la PT

Application coté cible (LM) :

- mêmes techniques
- sélection des données mono-lingue
- adaptation du LM

Sélection de données

Autre méthodes :

- sélectionner (intelligemment) des phrases sources
- traduire ces phrases par des humains
- augmenter les corpus d'apprentissage avec ces données

- traduire ces phrases par un SMT
- garder une partie de ces données (comment)
- augmenter les corpus d'apprentissage

Sélection de données (utilisation d'un LM)

- Apprendre un LMI sur les données *in-domain*
- Apprendre un LMO sur une partie (de même taille) les données *ou-of-domain*
- calculer *l'entropie croisée* pour chaque phrase *out-of-domain* avec LMI et LMO
- Les phrases sont ordonnées en fonction de la difference *d'entropie croisée*

$$LMI - LMO$$

Difference *d'entropie croisée*

Sélection mono-lingue *Moore and Lewis (2010)*

1. corpus *in-domain* I
2. corpus *out-of-domain* O
3. on sélectionne un sous ensemble $\hat{O} \in O$
   ($\hat{O}$ de même taille que I)
4. apprendre LM_I sur I
5. apprendre $LM_{\hat{O}}$ sur $\hat{O}$
6. pour chaque phrase de O
   $$\Rightarrow \text{calculer} : H_{LM_I}(o) - H_{LM_{\hat{O}}}(o)$$
7. garder les n première phrases dans le corpus d'apprentissage

Difference *d'entropie croisée*

L'entropie est une mesure de l'incertitude d'une variable aléatoire

L'*entropie croisée* est utilisée lorsqu'on ne connait pas la distribution de probabilité p qui a générée les données (le langage naturel). L'*entropie croisée* est un borne supérieure de l'entropie.

H est *l'entropie croisée* de ML normalisée par le longueur de chaque phrase :

$$H_{LM}(x) = -\sum_{i=1}^{|x|} \frac{1}{|x|} log \;\; p_{LM}(x_i|x_{i-1}) \text{ (lower is better)}$$

Cela permet de sélectionner les phrases similaire au corpus *in-domain* et non similaire du corpus *out-of-domain*

Difference de *cross-entropy*

Sélection bi-lingue (*Axelrod et al.* 2011)

- Deux corpus *in-domain* : I_{src} et I_{trg}
- Deux corpus *out-of-domain* : O_{src} et O_{trg}
- Deux sous partie de *out-of-domain* : $\hat{O}_{src} \in O_{src}$ $\hat{O}_{trg} \in O_{trg}$
- Quatre LM : $LM_{I_{src}}$, $LM_{I_{tgt}}$, $LM_{\hat{O}_{src}}$ et $LM_{\hat{O}_{tgt}}$
- pour chaque pair de phrase (s, t) calculer

$$H_{LM_{I_{src}}}(s) - H_{LM_{\hat{O}_{src}}}(s) + H_{LM_{I_{trg}}}(t) - H_{LM_{\hat{O}_{trg}}}(t)$$

(aussi lower is better)

Résultats

Domain Adaptation via Pseudo In-Domain Data Selection. Amittai Axelrod EMNLP 2011

Method	Sentences	Dev	Test
General	12m	42.62	40.51
Cross-Entropy	35k	39.77	40.66
Cross-Entropy	70k	40.61	**42.19**
Cross-Entropy	150k	42.73	**41.65**
Moore-Lewis	35k	36.86	40.08
Moore-Lewis	70k	40.33	39.07
Moore-Lewis	150k	41.40	40.17
bilingual M-L	35k	39.59	**42.31**
bilingual M-L	70k	40.84	**42.29**
bilingual M-L	150k	42.64	**42.22**

FIGURE: Score BLEU en fonction de la sélection données

Résultats

La sélection de données :

- Améliorer le score BLEU
  $\Rightarrow$ une meilleure traduction
- Avec moins de données $\Rightarrow$ des modèles plus compactes
- Données *out-of-domain* $\Rightarrow$ généralement disponible en mono-lingue $\Rightarrow$ peu disponible en bi-lingue (Langues sous-dotées)
- Besoin d'autres méthodes d'adaptation

Sommaire

Adaptation de la Table de traduction

Techniques :

- Apprendre plusieurs table de traduction
- Pondération des corpus
- création de données
- Ajout d'une ff domaine

Adaptation de la Table de traduction

Utilisation de plusieurs table de traduction :
$\Rightarrow$ Classification des données d'apprentissage en K classes

Durant le décodage :

- classification non supervisée des données de test
- On récupère des traduction de plusieurs TMs
- Pondération des traductions
- En fonction de la distance de la phrase à traduire avec la partie (LM sur la partie source de bitext)

Alternative à la classification non supervisée : utilisation des *gold labels* (si disponibles)

Adaptation de la Table de traduction

system	BLEU
in-domain	17.9
out-of-domain	14.8
all data (unweighted)	18.3
all data (weighted)	**18.8**

FIGURE: Exemple de traduction (domaine = alpinisme)

Adaptation de la Table de traduction

System	Satz
source:	Schön brav steigen wir zu Fuss zur Vallot-Hütte.
reference:	Bien sages, nous descendons à pied jusqu'au refuge Vallot.
in-domain system:	Il fait beau, nous montons gentiment à pied à la Vallot-Hütte.
Personal Translator 14:	Nous augmentons bien sagement à Fuss à la hutte Vallot.
Google Translate:	Nice bon on arrive à pied à la cabane Vallot.

FIGURE: Score BLEU DE–FR (domaine = alpinisme)

Adaptation de la Table de traduction

Dans la présence de plusieurs corpus : Comment apprendre la PT :

- concaténation (solution simple)
- Pondération des corpus :

  $\Rightarrow$ multiplier n fois le corpus x

  $\Rightarrow$ n est fonction de la distance de x du domaine

  $\Rightarrow$ Les séquences de mots proche de domain sont plus fréquent

Adaptation de la Table de traduction

Génération de données :

- R.I : chercher sur Internet des text (SRC) dans le domaine
- Traduire le text *du domaine* avec un système externe (GoogleT)
- Sélectionner une partie de text traduit (mesure de confiance)
- Ajouter la sélection aux données d'apprentissage
- Apprendre un nouveaux système

Adaptation de la Table de traduction

- Ajout d'une information d'appartenance au domaine (oui/non)
- Utilisation d'une fonction caractéristique pour le domaine
- Optimiser cette fonction avec *MERT*

Sommaire

Adaptation d'un modèles de langage

Deux LMs :

- Background LM : appris avec beaucoup de données
- Adaptation LM : peu de données mais *in-domain*

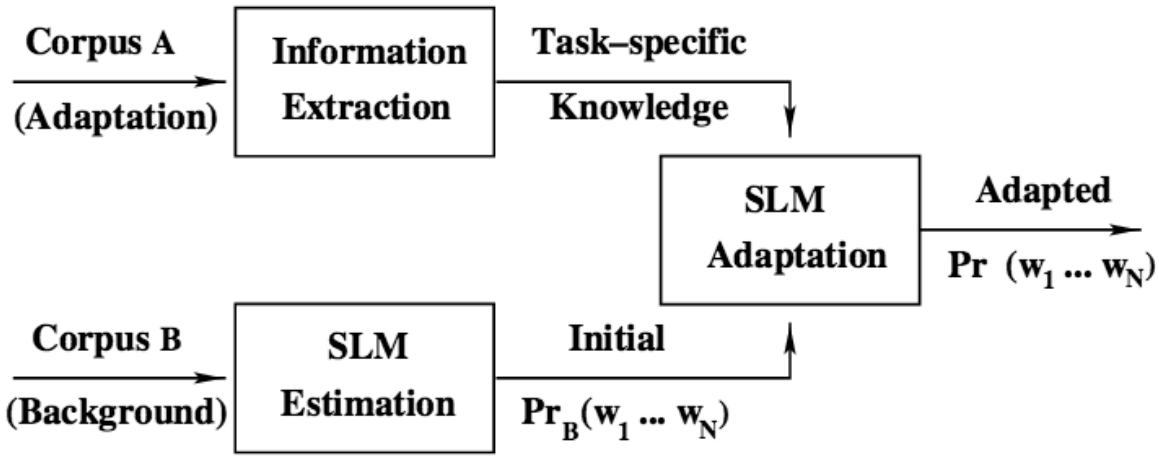

FIGURE: Adaptation d'un ML statistique

Adaptation d'un modèles de langage

- Utilisation de plusieurs LM dans le decodeur
  $\Rightarrow$ plusieurs LM comme différent ff dans Moses

- Interpolation de plusieurs LM (in et out-of-domain)
  $\Rightarrow \quad P(w_q|h_q) = \lambda P_a(w_q|h_q) + (1-\lambda)P(w_q|h_q)$
  λ : cœffecient d'interpolation

  $\Rightarrow$ Optimisation de coefficient d'interpolation sur un dev
- MAP adaptation : combinaison au niveau de comptage de fréquence de mot

Adaptation d'un modèles de langage

Adaptation d'un NNLM (la semaine prochaine)

Références

Y. Lu, J. Huang, and Q. Liu, Improving statistical machine translation performance by training data selection and optimization, *in Proc. of EMNLP-CoNLL, 2007, pp. 343–350*

A. Ittycheriah and S. Roukos, Direct translation model 2 *in Proc. of HLT/NAACL, 2007, pp. 57–64.*
Domain Adaptation via Pseudo In-Domain Data Selection. Amittai Axelrod EMNLP 2011