POLYTECH PARIS-SUD

Et5-Info

Module : Traitement Automatique des Langues

TP 2 – Traduction automatique statistique

Contexte :

La traduction automatique pour les langues morphologiquement riches nécessite un pré-traitement impliquant au minimum une analyse morpho-syntaxqiue des données d'apprentissage.

Traduction automatique statistique :

L'objectif de ce TP est l'installation du système de traduction automatique statistique Moses et son expérimentation sur des corpus parallèles en formes fléchies.

Travail demandé

Vous allez installer et expérimenter le système de traduction automatique statistique Moses.

1. Evaluation du système de traduction statistique Moses sur un corpus parallèle en formes fléchies

Informations sur Moses : http://www.statmt.org/moses/index.php?n=Main.HomePage

Guides d'installation :

- http://www.statmt.org/moses/?n=Moses.Baseline
- https://www2.statmt.org/moses/?n=Development.GetStarted

1. Installation

a. Installer les packages logiciels nécessaires à la compilation et l'installation de Moses (git, Boost, etc.).
b. Installer Moses.
c. Installer GIZA++ et MGIZA.
d. Installer IRSTLM.

2. Expérimentation

Réaliser des expérimentations sur le petit corpus parallèle joint pour le couple de langues anglais-français :

Train (Corpus d'apprentissage) :
Europarl_train_10k.en
Europarl_train_10k.fr

Dev (Corpus de développement) :

Europarl_dev_1k.en
Europarl_dev_1k.fr

Test (Corpus de test) :

Europarl_test_500.en
Europarl_test_500.fr

3. Evaluation

a. Réaliser des expérimentations sur deux corpus parallèles pour le couple de langues anglais-français. La taille (en nombre de phrases) de l'ensemble des corpus bilingues utilisés pour l'apprentissage et le développement du système de traduction Moses est décrite dans le tableau ci-dessous :

N° du run	**Apprentissage (nombre de phrases)**	**Tuning (nombre de phrases)**
1	100K (Europarl)	3,75K (Europarl)
2	100K+10K (Europarl+Emea)	3,75K (Europarl)

b. Evaluer Moses en utilisant le score BLEU. Il faudrait effectuer 2 runs avec, pour chacun d'entre eux, un test avec des données du domaine et un autre hors-domaine.

Notes :

1. Pour pouvoir comparer les différents résultats, je vous joins les données nécessaires pour cette évaluation.
2. Vous pouvez récupérer la totalité des corpus parallèles Europarl pour le domaine général et Emea pour le domaine médical à partir des liens http://opus.nlpl.eu/Europarl.php et http://opus.nlpl.eu/EMEA.php.

Corpus d'apprentissage :

- 100K (Europarl) : Europarl_train_100k.en, Europarl_train_100k.fr ➔ Il faut prendre les 100 000 premières phrases du corpus.
- 10K (Emea) : Emea_train_10k.en, Emea_train_10k.fr ➔ Il faut prendre les 10 000 premières phrases du corpus.

Corpus de développement :

- Europarl_dev_3750.en ➔ Il faut prendre 3750 phrases à partir du corpus Europarl en commençant par la phrase au rang 100 001
- Europarl_dev_3750.fr ➔ Il faut prendre 3750 phrases à partir du corpus Europarl en commençant par la phrase au rang 100 001

Corpus de test :

- Europarl : Europarl_test_500.en, Europarl_test_500.fr ➔ Il faut prendre 500 phrases à partir du corpus Europarl en commençant par la phrase au rang 103751
- Emea : Emea_test_500.en, Emea_test_500.fr ➔ Il faut prendre 500 phrases à partir du corpus Emea en commençant par la phrase au rang 10001

Remarques :

- L'idéal pour le corpus de test d'extraire de manière aléatoire un ensemble de 500 paires de phrases à partir du corpus Europarl comme un corpus correspondant au domaine et 500 autres paires de phrases à partir du corpus Emea comme un corpus hors-domaine.
- Vous pouvez utiliser la procédure train_test_split du module sklearn.model_selection pour créer les corpus d'apprentissage et de test via une partition au hasard.