Design Document

Description of the program architecture

Input :

A JSON file in a proper format, which contains French and English sentences which are translations of each other. Basically, it is a parallel corpora.

Working :

The program pre-processes the dataset to get sentences and later tokenizes these sentences. Then these sentence pairs are fed into the model which was implemented. This follows the entire procedure of expectation-maximization until the translation probabilities converge. It then prints the alignments obtained through the implemented model.

Following it, the same dataset is fed to IBM Model 1 from the nltk.translate module and its alignments calculated. This is repeated for IBM Model 2, too.

Then the alignments obtained from the implemented module is passed on as an argument to the phrase_extraction module, which extracts phrase pairs in the form of (english_phrase, french_phrase). We keep a count of the total number of pairs in which a particular french_phrase f has occurred , say count_1. Also for any particular english_phrase e, we check how many pairs (e,f) have occured, say count_2. Phrase score for (e,f) is count_2/count_1. Then all phrase pairs with non zero phrase scores are printed in descending order.

Output:

1) Alignments of coded model
2) Alignments of IBM Model 1
3) Alignments of IBM Model 2
4) Extracted phrase pairs in descending order of phrase translation scores.

Data Structures Used:

1) Lists have been used to store distinct english words, distinct french words, and after some processing, the original sentences are stored as lists as well. Alignments have been represented as lists of tuples as well. Finally, the phrase pairs and their scores are represented as a list of tuples.

2) Dictionaries for storing the translation probability corresponding to a pair (english_word, french_word) where the latter is the key, and the former is the value. Also dictionaries are used inside the expectation-maximization loop for storing total count of french words, and count(english_word|french_word). Furthermore, count of french phrases and count of pairs of (english_phrases, french_phrases) have also been stored in dictionaries.
3) Bitext has been used as input to the inbuilt IBM Models.

Results

For data1.json

Sentence Pairs :

["la maison", "the house"]
["la fleur", "the flower"]
["la maison bleu", "the blue house"]
["la fleur bleu", "the blue flower"]
["pomme bleu", "blue apple"]

Iterations to converge (threshold = 0.0001) = 178

Alignments from our implementation, IBM Model 1 and IBM Model 2 were the same. Presented below:

[(0, 0), (1, 1)]
[(0, 0), (1, 1)]
[(0, 0), (1, 2), (2, 1)]
[(0, 0), (1, 2), (2, 1)]
[(0, 1), (1, 0)]

Phrase translation scores :

(1.0, ('the house', 'la maison'))
(1.0, ('the flower', 'la fleur'))
(1.0, ('the blue house', 'la maison bleu'))
(1.0, ('the blue flower', 'la fleur bleu'))
(1.0, ('the', 'la'))

(1.0, ('house', 'maison'))
(1.0, ('flower', 'fleur'))
(1.0, ('blue house', 'maison bleu'))
(1.0, ('blue flower', 'fleur bleu'))
(1.0, ('blue apple', 'pomme bleu'))
(1.0, ('blue', 'bleu'))
(1.0, ('apple', 'pomme'))

For data2.json

Sentence Pairs :

["la maison est en inde", "the house is in india"]
["l' inde est verte", "india is green"]
["la fille est intelligent", "intelligent is the girl"]
["la fille verte est en inde", "in india is the green girl"]
["la fille a une chaise verte","the girl has a green chair"]
["sur la chaise verte est une fille","a girl is on the green chair"]
["dans l' inde est une chaise","a chair is in india"]
["la chaise est verte","the chair is green"]

Iterations to converge (threshold = 0.0001) = 507

Alignments from our implementation, IBM Model 1 and IBM Model 2 were the same. Presented below:

[(0, 0), (1, 1), (2, 2), (3, 3), (4, 4)]
[(0, 1), (1, 2), (2, 3)]
[(0, 3), (1, 2), (2, 0), (3, 1)]
[(0, 4), (1, 5), (2, 3), (3, 0), (4, 2), (5, 1)]
[(0, 0), (1, 1), (2, 2), (3, 3), (4, 5), (5, 4)]
[(0, 5), (1, 6), (2, 4), (3, 0), (4, 1), (5, 3), (6, 2)]
[(0, 4), (1, 5), (2, 3), (3, 0), (4, 2)]
[(0, 0), (1, 1), (2, 2), (3, 3)]

Phrase translation scores :

(1.0, ('the house is in india', 'la maison est en inde'))
(1.0, ('the house is in', 'la maison est en'))
(1.0, ('the house is', 'la maison est'))
(1.0, ('the house', 'la maison'))
(1.0, ('the green girl', 'la fille verte'))
(1.0, ('the green chair', 'la chaise verte'))
(1.0, ('the girl has a green chair', 'la fille a une chaise verte'))
(1.0, ('the girl has a', 'la fille a une'))
(1.0, ('the girl has', 'la fille a'))
(1.0, ('the girl', 'la fille'))
(1.0, ('the chair is green', 'la chaise est verte'))
(1.0, ('the chair is', 'la chaise est'))
(1.0, ('the chair', 'la chaise'))
(1.0, ('the', 'la'))
(1.0, ('on the green chair', 'sur la chaise verte'))
(1.0, ('on the', 'sur la'))
(1.0, ('on', 'sur'))
(1.0, ('is the green girl', 'la fille verte est'))
(1.0, ('is the girl', 'la fille est'))
(1.0, ('is on the green chair', 'sur la chaise verte est'))
(1.0, ('is in india', "dans l' inde est"))
(1.0, ('is in', 'est en'))
(1.0, ('is green', 'est verte'))
(1.0, ('is', 'est'))
(1.0, ('intelligent is the girl', 'la fille est intelligent'))
(1.0, ('intelligent is', 'est intelligent'))
(1.0, ('intelligent', 'intelligent'))
(1.0, ('india is green', "l' inde est verte"))
(1.0, ('india is green', 'inde est verte'))
(1.0, ('india is', "l' inde est"))
(1.0, ('india is', 'inde est'))
(1.0, ('india', "l' inde"))
(1.0, ('india', 'inde'))
(1.0, ('in india is the green girl', 'la fille verte est en inde'))
(1.0, ('in india', 'en inde'))
(1.0, ('in india', "dans l' inde"))
(1.0, ('in', 'en'))

(1.0, ('in', "dans l'"))
(1.0, ('in', 'dans'))
(1.0, ('house is in india', 'maison est en inde'))
(1.0, ('house is in', 'maison est en'))
(1.0, ('house is', 'maison est'))
(1.0, ('house', 'maison'))
(1.0, ('has a green chair', 'a une chaise verte'))
(1.0, ('has a', 'a une'))
(1.0, ('has', 'a'))
(1.0, ('green girl', 'fille verte'))
(1.0, ('green chair', 'chaise verte'))
(1.0, ('green', 'verte'))
(1.0, ('girl has a green chair', 'fille a une chaise verte'))
(1.0, ('girl has a', 'fille a une'))
(1.0, ('girl has', 'fille a'))
(1.0, ('girl', 'fille'))
(1.0, ('chair is green', 'chaise est verte'))
(1.0, ('chair is', 'chaise est'))
(1.0, ('chair', 'chaise'))
(1.0, ('a green chair', 'une chaise verte'))
(1.0, ('a girl is on the green chair', 'sur la chaise verte est une fille'))
(1.0, ('a girl is', 'est une fille'))
(1.0, ('a girl', 'une fille'))
(1.0, ('a chair is in india', "dans l' inde est une chaise"))
(1.0, ('a chair is', 'est une chaise'))
(1.0, ('a chair', 'une chaise'))
(1.0, ('a', 'une'))
(0.5, ('is in india', 'est en inde'))
(0.5, ('in india is', 'est en inde'))

For our corpus (data3.json)

Sentence Pairs :

["la fille est en france", "the girl is in france"]
["paris est une ville en france", "paris is a city in france"]
["la fille est belle", "the girl is beautiful"]
["paris est une belle ville", "paris is a beautiful city"]
["la fille est a la eglise","the girl is in church"]
["une eglise est en paris","a church is in paris"]
["la eglise est belle","the church is beautiful"]
["la france a une belle eglise","france has a beautiful church"]

Iterations to converge (threshold = 0.0001) = 235

Alignments from our implementation, IBM Model 1 and IBM Model 2 were the same. Presented below:

[(0, 0), (1, 1), (2, 2), (3, 3), (4, 4)]
[(0, 0), (1, 1), (2, 2), (3, 3), (4, 4), (5, 5)]
[(0, 0), (1, 1), (2, 2), (3, 3)]
[(0, 0), (1, 1), (2, 2), (3, 3), (4, 4)]
[(0, 0), (1, 1), (2, 2), (3, 3), (4, 5)]
[(0, 0), (1, 1), (2, 2), (3, 3), (4, 4)]
[(0, 0), (1, 1), (2, 2), (3, 3)]
[(0, 1), (1, 2), (2, 3), (3, 4), (4, 5)]

Expected alignments :

[(0, 0), (1, 1), (2, 2), (3, 3), (4, 4)]
[(0, 0), (1, 1), (2, 2), (3, 3), (4, 4), (5, 5)]
[(0, 0), (1, 1), (2, 2), (3, 3)]
[(0, 0), (1, 1), (2, 2), (3, 3), (4, 4)]
[(0, 0), (1, 1), (2, 2), (3, 3), (4, 5)]
[(0, 0), (1, 1), (2, 2), (3, 3), (4, 4)]
[(0, 0), (1, 1), (2, 2), (3, 3)]
[(0, 1), (1, 2), (2, 3), (3, 4), (4, 5)]

Phrase translation scores :

(1.0, ('the girl is in france', 'la fille est en france'))
(1.0, ('the girl is in church', 'la fille est a la eglise'))
(1.0, ('the girl is in church', 'fille est a la eglise'))
(1.0, ('the girl is in', 'la fille est en'))
(1.0, ('the girl is in', 'la fille est a la'))
(1.0, ('the girl is in', 'fille est a la'))
(1.0, ('the girl is beautiful', 'la fille est belle'))
(1.0, ('the church is beautiful', 'la eglise est belle'))
(1.0, ('the church is', 'la eglise est'))
(1.0, ('the church', 'la eglise'))
(1.0, ('the', 'la'))
(1.0, ('paris is a city in france', 'paris est une ville en france'))
(1.0, ('paris is a city in', 'paris est une ville en'))
(1.0, ('paris is a city', 'paris est une ville'))
(1.0, ('paris is a beautiful city', 'paris est une belle ville'))
(1.0, ('paris is a beautiful', 'paris est une belle'))
(1.0, ('paris is a', 'paris est une'))
(1.0, ('paris is', 'paris est'))
(1.0, ('paris', 'paris'))
(1.0, ('is in paris', 'est en paris'))
(1.0, ('is in france', 'est en france'))
(1.0, ('is in', 'est en'))
(1.0, ('is in', 'est a'))
(1.0, ('is beautiful', 'est belle'))
(1.0, ('is a city in france', 'est une ville en france'))
(1.0, ('is a city in', 'est une ville en'))
(1.0, ('is a city', 'est une ville'))
(1.0, ('is a beautiful city', 'est une belle ville'))
(1.0, ('is a beautiful', 'est une belle'))
(1.0, ('is a', 'est une'))
(1.0, ('is', 'est'))
(1.0, ('in paris', 'en paris'))
(1.0, ('in france', 'en france'))
(1.0, ('in', 'en'))
(1.0, ('has a beautiful church', 'a une belle eglise'))
(1.0, ('has a beautiful', 'a une belle'))
(1.0, ('has a', 'a une'))

(1.0, ('girl is in france', 'fille est en france'))
(1.0, ('girl is in', 'la fille est a'))
(1.0, ('girl is in', 'fille est en'))
(1.0, ('girl is in', 'fille est a'))
(1.0, ('girl is beautiful', 'fille est belle'))
(1.0, ('girl is', 'fille est'))
(1.0, ('girl', 'fille'))
(1.0, ('france has a beautiful church', 'la france a une belle eglise'))
(1.0, ('france has a beautiful church', 'france a une belle eglise'))
(1.0, ('france has a beautiful', 'la france a une belle'))
(1.0, ('france has a beautiful', 'france a une belle'))
(1.0, ('france has a', 'la france a une'))
(1.0, ('france has a', 'france a une'))
(1.0, ('france has', 'la france a'))
(1.0, ('france has', 'france a'))
(1.0, ('france', 'la france'))
(1.0, ('france', 'france'))
(1.0, ('city in france', 'ville en france'))
(1.0, ('city in', 'ville en'))
(1.0, ('city', 'ville'))
(1.0, ('church is in paris', 'eglise est en paris'))
(1.0, ('church is in', 'eglise est en'))
(1.0, ('church is beautiful', 'eglise est belle'))
(1.0, ('church is', 'eglise est'))
(1.0, ('church', 'eglise'))
(1.0, ('beautiful city', 'belle ville'))
(1.0, ('beautiful church', 'belle eglise'))
(1.0, ('beautiful', 'belle'))
(1.0, ('a city in france', 'une ville en france'))
(1.0, ('a city in', 'une ville en'))
(1.0, ('a city', 'une ville'))
(1.0, ('a church is in paris', 'une eglise est en paris'))
(1.0, ('a church is in', 'une eglise est en'))
(1.0, ('a church is', 'une eglise est'))
(1.0, ('a church', 'une eglise'))
(1.0, ('a beautiful city', 'une belle ville'))
(1.0, ('a beautiful church', 'une belle eglise'))
(1.0, ('a beautiful', 'une belle'))
(1.0, ('a', 'une'))

(0.6666666666666666, ('the girl is', 'la fille est'))
(0.6666666666666666, ('the girl', 'la fille'))
(0.5, ('in', 'a'))
(0.5, ('has', 'a'))
(0.3333333333333333, ('girl is', 'la fille est'))
(0.3333333333333333, ('girl', 'la fille'))

Discussion of Results

1) The alignments from all the three modules were same because the data corpora were not large enough for any observable difference.
2) Most of the pairs of phrases will have the phrase scores of 1.0 (i.e, a perfect score), again because the datasets are small. Most French phrases would have a single English phrase in translation within the dataset.