Unsupervised extraction of bilingual lexicon

Caio Corro, Guillaume Wisniewski

1 Introduction

The goal of this lab exercise is to understand that we can learn something from data with minimal supervision, that is unsupervised (or weakly supervised) learning can actually work.

This lab aims at extracting automatically a bilingual lexicon from a corpus. A bilingual lexicon is a dictionary that associates a word of a *source language* to the list of its possible translations in a *target language*. Table 1 shows an excerpt of a French-English lexicon.

cuisine	cooks, cook, cooking, kitchen, food...
comprendre	understand, understood, understanding, realize, ...
être	human, being, be, ...
économie	free-market, Economy, economics, market, economy, ...

Table 1 – Excerpt of a French-English lexicon.

The extraction method we propose is *unsupervised* : the lexicon will be extracted from a corpus of parallel sentences (i.e. a French sentence and its translation in English) without any information about the expected output. Table 2 shows an example of a parallel corpus. This method relies on the *distributional hypothesis* : the meaning of a word can be deduced from the context in which it appears. In the bilingual case, this hypothesis can be formulated as follows : a French word and an English word which often *co-occurs* (i.e. appear in a sentence and its translation) are likely to be translations. For instance, considering the parallel corpus of Table 2, it can naturally be deduced that *cuisine* is translated either by *cooking* or by *kitchen* as these two words are appearing in all translations of a sentence containing *cuisine*.

1. Considering the French-Greek corpus described in Table 3 can you guess the translation in Greek of `Elli`, `est` and `maison` ?

2 Naive approach

You will find on the lecture website two files containing the novel *Voyage au centre de la Terre* in French and English. The text has been pre-processed to :

— align sentences : the *i*-th line of the French file is the translation of the *i*-th line of the English file ;

Living on my own, I really miss my Mom’s **cooking**. Vivant seul, la **cuisine** de ma mère me manque.
She left the **kitchen** with the kettle boiling. Elle quitta la **cuisine** avec la bouilloire.
Is there any coffee in the **kitchen**? Y a-t-il encore du café dans la **cuisine** ?
Cooking runs in my family. La **cuisine** c’est de famille.
Both boys and girls should take **cooking** class in school. Garçons et filles devraient suivre des cours de **cuisine** à l’école.

Table 2 – Example of a English-French parallel corpus

Πάω στο σπίτι μας. Je vais chez nous (*lit.* dans notre maison).
Το σπίτι μου είναι μεγάλο. Ma maison est grande
Το σπίτι της Έλλης είναι κοντά στην παραλία. La maison d’Elli est à côté de la plage.
Με λένε Έλλη. Je m’appelle Elli.
Ένα σπίτι του χωριού κάηκε Une maison du village a brûlé.
Αγαπώ την Έλλη. J’aime Elli.

Table 3 – Exemple d’un corpus parallèle grec–français

— *tokenized* into words : tokenization consists in ensuring that there is a space before and after each words. For instance `"à_l’école."` is transformed into `"à_l’_école_."` and "100 000" into "100000".

Extracting a lexicon from a parallel corpus relies on the construction of a co-occurrence table that can be modeled by a dictionary of dictionaries that maps a French word (the first key) to a dictionary the keys of which are all the English words co-occurring with the French word ; the values of the second dictionary are the number of sentences in which a French and an English word both appear. For instance, if the corpus is made of the two sentences in Table 4, the extracted table will be :

```
{('chat', 'and'): 1,
 ('chat', 'cat'): 1,
 ('chat', 'dog'): 1,
 ('chat', 'the'): 1,
 ('chien', 'and'): 1,
 ('chien', 'cat'): 1,
 ('chien', 'dog'): 1,
 ('chien', 'the'): 1,
 ('et', 'and'): 2,
 ('et', 'calf'): 1,
 ('et', 'cat'): 1,
 ('et', 'cow'): 1,
 ('et', 'dog'): 1,
 ('et', 'the'): 2,
 ('la', 'and'): 1,
 ('la', 'calf'): 1,
 ('la', 'cow'): 1,
 ('la', 'the'): 1,
 ('le', 'and'): 2,
 ('le', 'calf'): 1,
 ('le', 'cat'): 1,
 ('le', 'cow'): 1,
 ('le', 'dog'): 1,
 ('le', 'the'): 2,
 ('vache', 'and'): 1,
 ('vache', 'calf'): 1,
 ('vache', 'cow'): 1,
 ('vache', 'the'): 1,
 ('veau', 'and'): 1,
 ('veau', 'calf'): 1,
 ('veau', 'cow'): 1,
 ('veau', 'the'): 1}
```

2. Why do we have to tokenized the novel ?

doc n^o 1	doc n^o 2
la vache et le veau	the cow and the calf
le chien et le chat	the dog and the cat

TABLE 4 – Example of two parallel sentences.

3. Write a method `count_sentences_with_word` that takes as input an open file and returns a dictionary that associates each word of the file to the number of sentences it appears in. **Be careful** : we are interested in the number of sentence a word is appearing in and not the number of times a word is occurring : if a word appears twice in a sentence, it should only be counted once.
4. Write a method `build_cooc_table` that takes as input two open files describing a parallel corpus and returns the co-occurence table. Again : it is important to count words that are repeated in a sentence only once.
5. Dump the content of the co-occurence table by decreasing frequency and analyze the result. Is this lexicon “good” ?

3 Statistical Tests

The method described in the previous sentence can only be used if we can identify ‘meaningful’ associations between a word and its translation from spurious associations are, for instance, due to frequent words.

Making this kind of distinction is the goal of significance tests. In the rest of this lab, we propose to consider one of this test, the *likelihood ratio* to filter the co-occurence table built in the previous Section and keep only meaningful associations. This test assess the significativity of the association between a word a and a word b from its contingency table that describes :

— $n(a, b)$ the number of sentence paris in which a and b co-occur ;
— $n(\neg a, b)$ the number of sentence pairs in which b appears but not a ;
— $n(a, \neg b)$ the number of sentence pairs in which a appears but not b ;
— $n(\neg a, \neg b)$ in which neither a nor b appear ;

Table 5 summarizes these notations.

	# sentences with a	# sentences without a	
# sentences with b	$n(a, b)$	$n(\neg a, b)$	$n(b)$
# sentences without b	$n(a, \neg b)$	$n(\neg a, \neg b)$	$n(\neg b)$
	$n(a)$	$n(\neg a)$	N

TABLE 5 – Table de contingence

The *likelihood ratio* is defined by :

$$G^2 = 2 \cdot N \left[\sum_{a? \in \{a, \neg a\}} \sum_{b? \in \{b, \neg b\}} p(a? \text{ and } b?) \cdot \log \frac{p(a? \text{ and } b?)}{p(a?) \cdot p(b?)} \right] \tag{1}$$

where $p(x?)$ is the frequency of the event $x?$ (the ratio between the number of times where $x?$ is true and the number of times where $x?$ is true *or* false). For instance, $p(a) = \frac{n(a)}{N}$ and $p(a, \neg b) = \frac{n(a, \neg b)}{N}$.

6. Let us denote $n(a)$ (resp. $n(b)$) the number of sentences in which a (resp. b) appears and N the number of sentences. How $n(\neg a, \neg b)$, $n(\neg a, b)$ and $n(a, \neg b)$ can be defined with respect to $n(a)$, $n(b)$, $n(a, b)$ and N.
7. Write a function that computes the *likelihood ratio* of a pair (French word, English word). The parameters of this method will be :
   - the number of sentences in the corpus ;
   - the number of sentences in which the French word appears ;
   - the number of sentences in which the English word appears ;
   - the number of parallel sentences in which both the English and French words appear ;
8. Write a method that computes the likelihood ratio of all pair of words. This method will return a dictionary mapping pair of words to their likelihood ratio. Pairs for which the likelihood is infinite must be discarded.
9. Dump all the pairs of words sorted by decreasing likelihood ratio. What can you conclude ?