

Deep Neural Models for Bilingual Lexicon Extraction

Jon Gauthier and Arthur Tsang

28 May 2014

Introduction

Neural networks and translation

The task: Bilingual lexicon extraction

Standard lexical learning

Minimally-supervised approaches

Introduction

Neural networks and translation

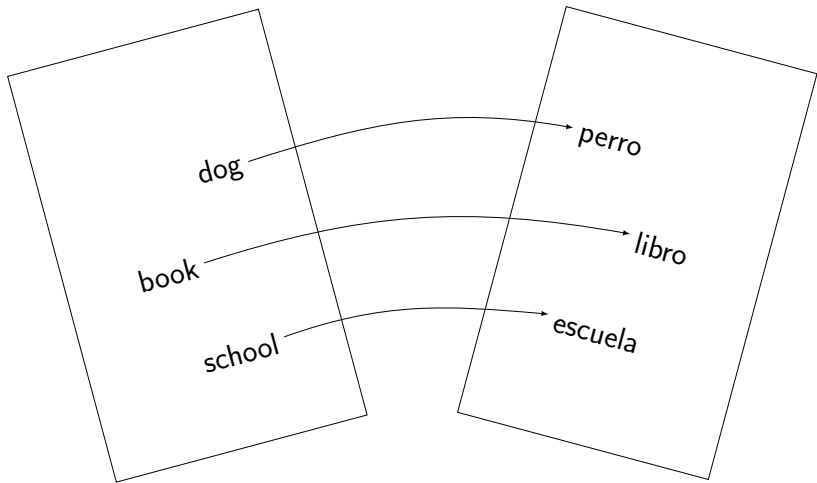
The task: Bilingual lexicon extraction

Standard lexical learning

Minimally-supervised approaches

How about this one?

$$f(\text{English}) = \text{Spanish}$$



Introduction

Neural networks and translation

The task: Bilingual lexicon extraction

Standard lexical learning

Minimally-supervised approaches

Bilingual lexicon extraction

Given multilingual corpora (composed of a *source language* and *target language*),

Bilingual lexicon extraction

Given multilingual corpora (composed of a *source language* and *target language*),

learn **translation mappings** of the form

Bilingual lexicon extraction

Given multilingual corpora (composed of a *source language* and *target language*),

learn **translation mappings** of the form

(*source language word*, *target language word*)

Standard lexical learning

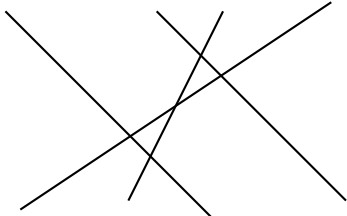
the cat is here

aquí está el gato

Standard lexical learning

the cat is here

aquí está el gato



But we need parallel corpora for this..

Soy siervo del Fuego Secreto,
administrador de la llama de
Anor. ¡Tu fuego oscuro es en
vano, llama de Udûn! Regresa a
las sombras. ¡No puedes pasar!
¡Corred, insensatos!

I am a servant of the secret fire,
wielder of the flame of Anor. The
dark fire will not avail you, flame
of Udûn! Go back to the
Shadow. You shall not pass! Fly,
you fools!

Non-parallel (comparable) corpora

Las **letras**, como elementos de los **alfabetos**, tienen un orden prescrito. Esto se conoce como “orden **alfabético**,” aunque la clasificación **alfabética** es la ciencia dedicada a la tarea compleja de ordenar y clasificar las **letras** en los diferentes **idiomas**.

A **letter** is a grapheme (written character) in an **alphabetic** system of writing, such as the **alphabet** of the Greek **language** and its descendants.