Empowering Citizens. Smarter Societies.

Insight
Centre for Data Analytics

Inferring Translation Candidates for Multilingual Dictionary Generation with Multi-Way Neural Machine Translation

Mihael Arcan, Daniel Torregrosa*, Sina Ahmadi* and John P. McCrae

A World Leading SFI Research Centre

This publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289, co-funded by the European Regional Development Fund, and the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015, ELEXIS - European Lexical Infrastructure.

Insight
Centre for Data Analytics

Introduction

Neural machine translation

Results

Dictionary data

Conclusion

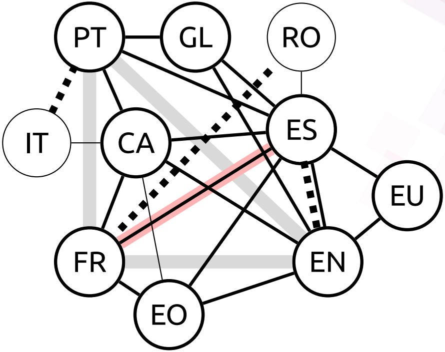
Motivation

- Knowledge bases are useful for many applications, but available in few languages
- The creation and curation of knowledge bases is expensive
- Hence, few or no knowledge bases in most languages
- Can we use machine translation to translate knowledge?

Overview

- Multi-way neural machine translation without the targeted direction
- Continuous training with a small curated dictionary
- Discovery of new bilingual dictionary entries

Targeted languages



Introduction

Neural machine translation

Results

Dictionary data

Conclusion

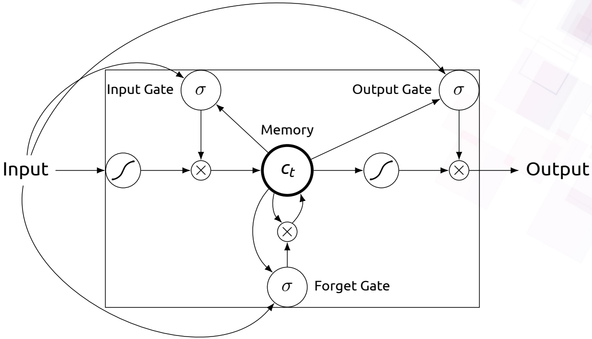
Machine translation before 2014

- Rule-based machine translation
  - Humans write rules
  - Highly customisable
  - High maintenance cost
- Phrase-based statistical machine translation
  - Learns from parallel corpus
  - Less control on the translations

Word embeddings

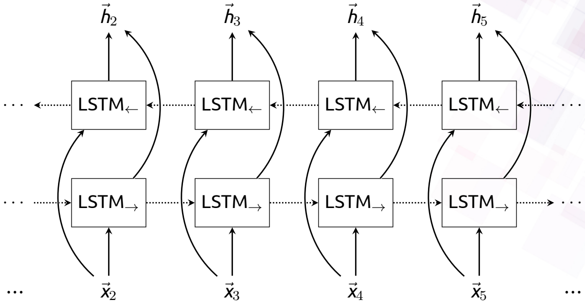
- Fixed size numerical representation for words
- From one-hot space (one dimension per different word) to embedding space
- The embedding vector represents the context where the word appears

Long-short term memory



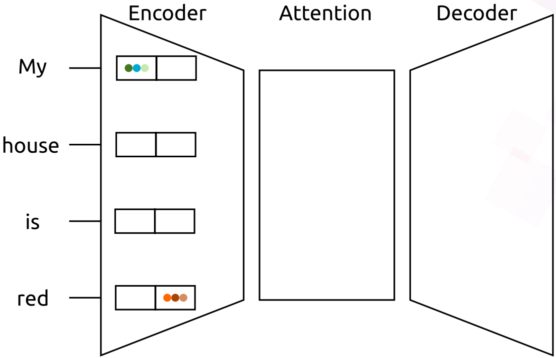
Based on `tex.stackexchange.com/questions/332747/how-to-draw-a-diagram-of-long-short-term-memory`

Bi-directional LSTM

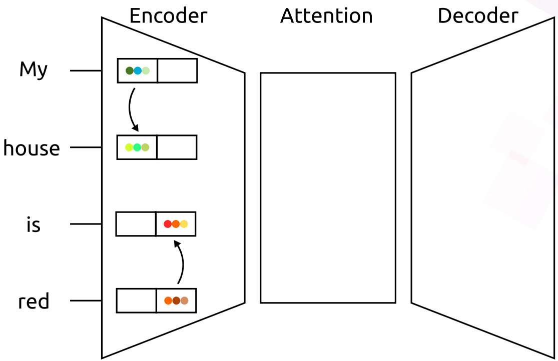


Based on `github.com/PetarV-/TikZ`

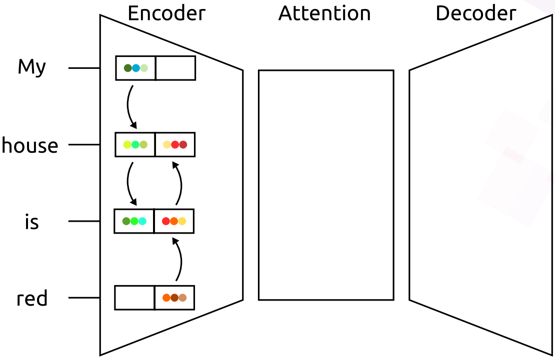
Neural machine translation



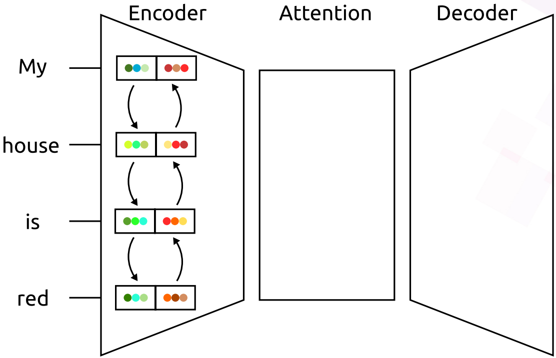
Neural machine translation



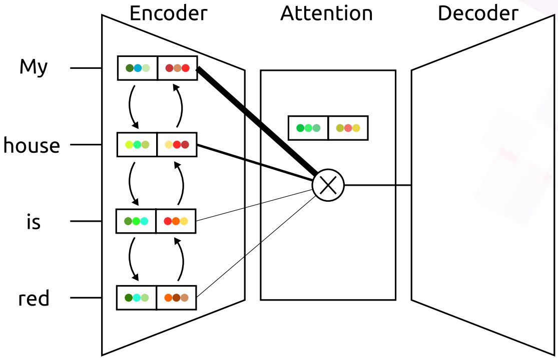
Neural machine translation



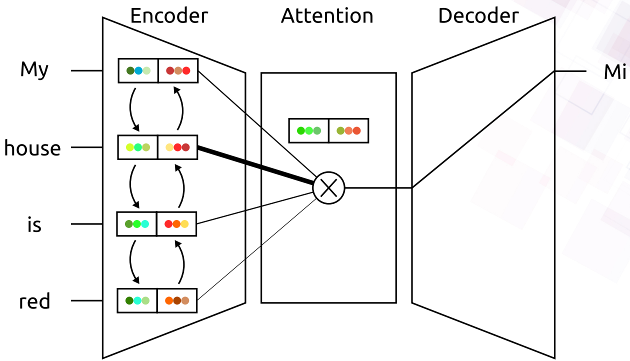
Neural machine translation



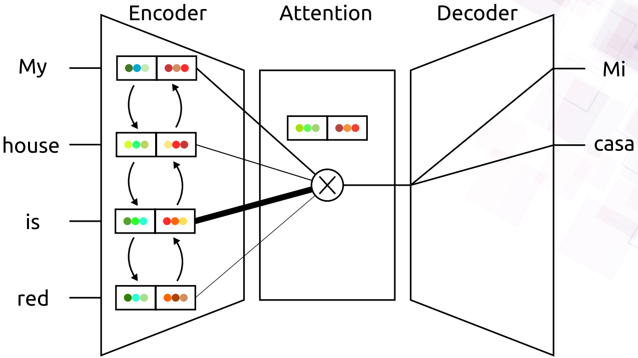
Neural machine translation



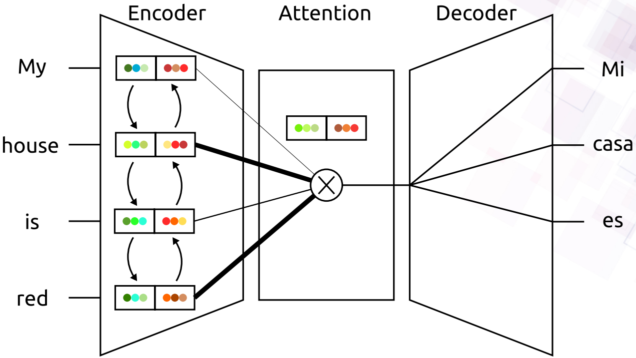
Neural machine translation



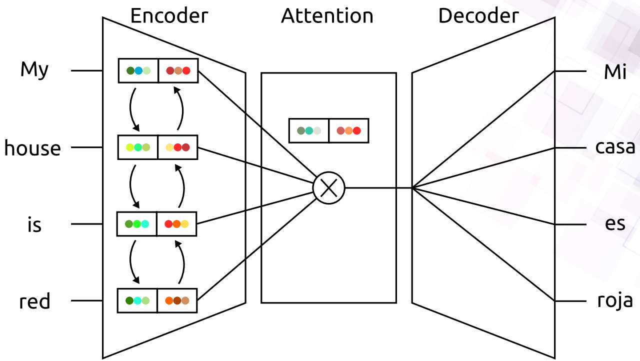
Neural machine translation



Neural machine translation



Neural machine translation



Subword units

- One-hot vocabulary space has to be limited due to performance issues
- This generates a lot of out-of-vocabulary entries
- To minimize the effect, we use subword units instead of words

Byte pair encoding

- BPE is a compression technique
- It starts with all the different characters in the corpus
- The most frequent character combination is selected as a BPE operation
- This is repeated until the desired number of BPE is reached
- The final size of the vocabulary is the number of BPE operations + the alphabet

Byte pair encoding example

low lower big bigger

Byte pair encoding example

l o w _ l o w e r _ b i g _ b i g g e r

Byte pair encoding example

l o w _ **l o** w e r _ b i g _ b i g g e r

Byte pair encoding example

l■o w _ l■o w e r _ **b i** g _ **b i** g g e r

Byte pair encoding example

l▪o w _ **l▪o w** e r _ b▪i g _ b▪i g g e r

Byte pair encoding II

Present	bebo bebes bebe	bebemos bebéis beben	Conditional	bebería beberías bebería	beberíamos beberíais beberían
Preterit	bebí bebiste bebió	bebimos bebisteis bebieron	Future	beberé beberás beberá	beberemos beberéis beberán
Imperfect	bebía bebías bebía	bebíamos bebíais bebían			

Byte pair encoding II

Present	**beb**o **beb**es **beb**e	**beb**emos **beb**éis **beb**en	Conditional	**beber**ía **beber**ías **beber**ía	**beber**íamos **beber**íais **beber**ían
Preterit	**beb**í **beb**iste **beb**ió	**beb**imos **beb**isteis **beb**ieron	Future	**beber**é **beber**ás **beber**á	**beber**emos **beber**éis **beber**án
Imperfect	**beb**ía **beb**ías **beb**ía	**beb**íamos **beb**íais **beb**ían			

Multi-way model

- The model receives corpus in several different languages both for source and target sentences
- Each input sentence is annotated with the source language and the requested target language
- In our case, Spanish-English, French-Romanian and Italian-Portuguese

Continuous training

- After training, the network is seldom able to produce text in the requested language other than the training one
- For example, if requested to translate Spanish to French, it will generate English
- We continue the training with a small corpus of sentences

Dictionary data

We used three different dictionaries to continue training the system

- Spanish to French Apertium dictionary (paper)

- Spanish-French, Spanish-Portuguese and French-Portuguese dictionaries generated from Apertium data (task)
  - By following a cycle-based approach

  - By following a path-based approach

Part of speech

- The NMT models were trained without part of speech (POS) data
- To assign POS, we use monolingual dictionaries automatically extracted from Wiktionary
- If
  - the source word is in the source-language dictionary; and
  - the target word is in the target-language dictionary; and
  - they have one or more POS tags in common,
- generate one entry per shared POS

Introduction

Neural machine translation

Results

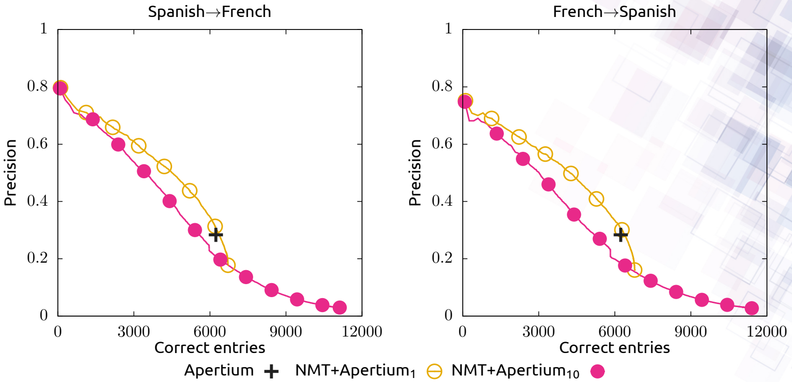
Dictionary data

Conclusion

Evaluation

- We used a dictionary automatically extracted from Wiktionary as gold standard
- For those systems that have confidence intervals, we calculate the precision and recall for all possible thresholds

Results (paper)



Introduction

Neural machine translation

Results

Dictionary data

Conclusion

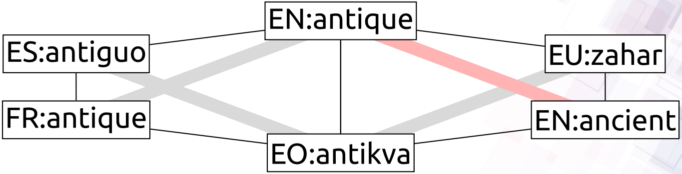
Graph-based approaches

Basic idea: Retrieve translations based on the graph of languages
Two definitions:

- Language graph refers to the Apertium dictionary graph

- Translation graph refers to a graph where vertices represent a word and edges represent the translations in other languages.

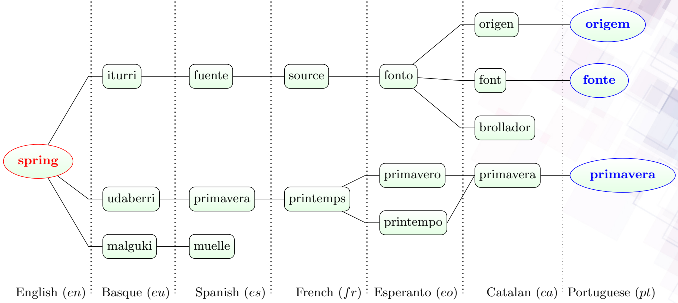
Cycle-based approach



Apertium translations (black lines) in English (EN), French (FR), Basque (EU) and Esperanto (EO), and discovered possible translations (gray lines) and synonyms (red lines).

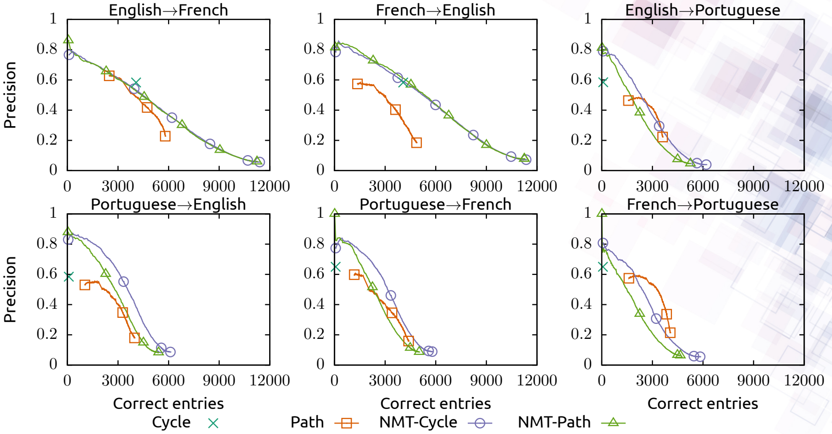
Path-based approach

Traverse all simple paths using **pivot-oriented inference**



(Task) Weight translations w.r.t. frequency and path length

Results (task, Wiktionary reference)



Introduction

Neural machine translation

Results

Dictionary data

Conclusion

Conclusion

- Using neural machine translation with
  - Existing bilingual knowledge (Paper)
  - Discovered bilingual knowledge (Task)
- to generate new dictionaries.