Egyptian Hieroglyphs and Machine Translation

Jiageng Liu

CS 269: Machine Learning for Natural Language Processing

UCLA, November 2017

In this blog, I present a recent paper on unsupervised translation. I also share my own thoughts on how the paper's main idea is connected to the deciphering of Egyptian Hieroglyphs, and suggest some future work.

Introduction

Egyptian Hieroglyphs

Translation has challenged scholars since ancient times. Egyptian hieroglyphs amazed classic scholars with their vivid drawings but equally confused them due to their lack of understanding. The language has died out thousands of years ago, and no current writing systems resemble it in any aspect. If we only take each glyph for its pictographic meaning, they do not group into any sentence that makes sense. Horapollo from ancient Greece attempted to make sense of them in his *Hieroglyphica* around the 4th century AD. He claimed that some pictures are abstract concepts, but his explanations were far from convincing. For example, he believed that hares represent opening because this animal always has their eyes open. The number 1095 represents dumbness, because if in three years (1095 days) a child cannot learn to speak, then it must be dumb. The funny explanation illustrates how hard it could be to translate a language without parallel reference, or, in modern computer science terminology, to do **unsupervised translation**.

Egyptian hieroglyphs

Unsupervised Translation

Most modern languages spoken today have detailed grammar books and dictionaries to help any human to learn language translation. The same holds for phrase-based machine translation. Programmers used to dump millions of phrases and syntax rules into the computer program, but the programs still do not perform very well.

Since 2014, however, **neural machine translation (NMT)** has completely changed the field. Researchers have proposed powerful Long-Short Term Memory (LSTM) networks along with recent improvements like the seq2seq model and the attention models. With the help of these neural networks, recent systems have made significant progress to generate sentences that are both human-readable and grasp the meaning of original texts. The improvement is so huge that in two years, Google and Microsoft have fully adopted the technology into their translation products.

Big improvement to Google Translate in 2017, from Stanford CS224n

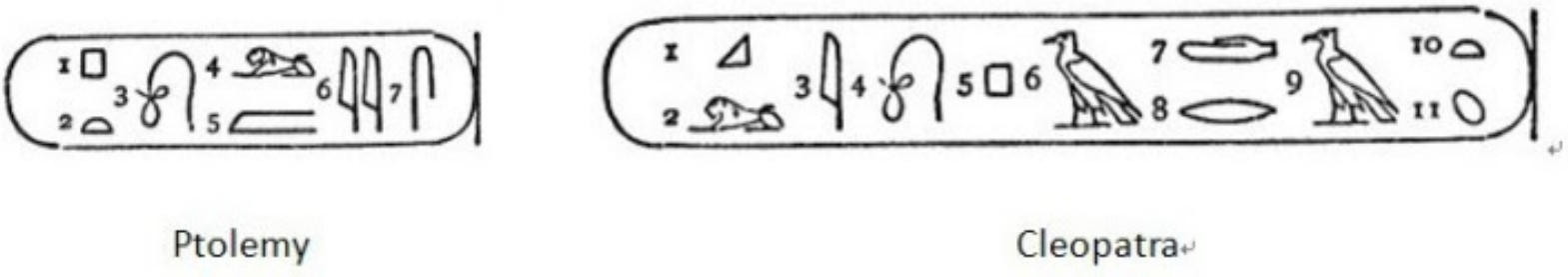
Neural networks also excited many programmers since it does not ask for any hard-coded rules, so there is no more tedious work like translating dictionaries into code. Instead, neural networks are known for their thirst for a huge amount of data. NMT networks, in particular, require gigabytes of text data in the form of **parallel corpora**, where the same texts in multiple languages are put side-by-side to teach the network how to translate.

The idea sounds great, but parallel corpora are very rare. Computer vision researchers often hire people to do simple tasks like image recognition to generate datasets. But it's way more expensive to hire foreign language experts to produce parallel corpora. Scientists have hence relied on multi-lingual official documents from European Parliament and United Nations. Yet the dependence makes NMT perform much worse on short texts. After all, on the dataset of laws and policies, the system is never trained to translate "where is the restroom," which ironically is what more people care about when they travel abroad. In a word, we want to translate all kinds of texts, and we want more data at a minimal cost.

Can we train NMT without parallel corpora data, or equivalently, can we train NMT to do unsupervised machine translation? Today, I am going to present a paper titled **Unsupervised Neural Machine Translation** published weeks ago. The work was done by Mikel Artetxe et al. at the University of Basque County and Kyunghyun Cho at New York University.

Cross-Lingual Tokens

We talk about the history of hieroglyphs for a reason. Unsupervised NMT shares a common key idea with the deciphering of that mysterious writing system, which owes credit to French linguist Jean-François Champollion. Champollion was the first to recognize that many glyphs do not carry graphical meanings, but are phonetic glyphs that stand for pronunciation. He also recalled that some foreign pharaohs ruled Egypt after Alexander the Great's conquest, so he successfully identified the names of two foreign pharaohs (Ptolemy and Cleopatra), whose names were certainly spelled in phonetic glyphs. Champollion used the consonants he deciphered from the names and combined them with the language's surviving traits in today's Egyptian languages, and finally understood most of the hieroglyphs.



First Cross-Lingual Tokens Identified by Champollion

The so-called "**cross-lingual tokens**" that Champollion identified are also the key to today's unsupervised NMT. Without parallel corpora, anything that connects the speakers' experiences to ours can serve as a clue. Champollion used history, and today we can use Arabic numerals.

Indeed, imagine you are reading the texts shown below. Even if you do not know what languages they are, you can still realize that the underlined texts above are dates, and those below refer to some period of time. You can further guess that the unknown characters above mean year, month, and day, and the texts below contain some person's name. We can do inference based on a basic assumption: Arabic numerals mean the same across modern languages. Based on our experiences with our mother tongues, we know that the texts surrounding them are limited to a few meanings, such as dates, units, or events.

· 다짐하면서 1948년 7월 12일에
정된 헌법을 이제 국회의 의결을
정한다. 1987년 10월 29일.

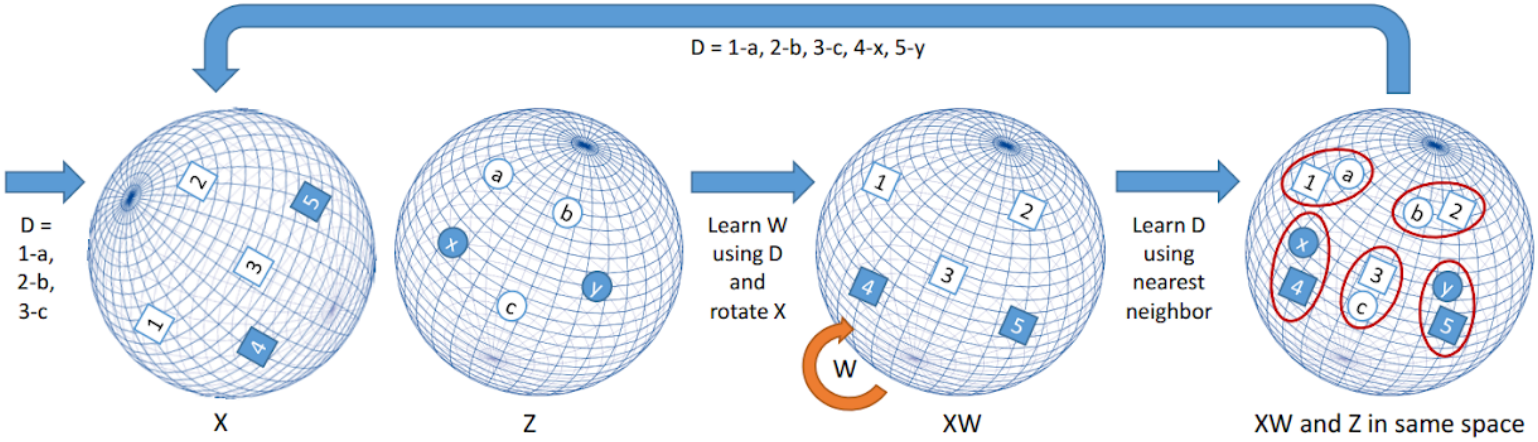
الملك عبد الله بن عبد العزيز (2015 - 2005) (يمين)، وسلمان بن عبد العزيز (الملك الحالي للبلاد) (يسار).

Modern Cross-Lingual Tokens, from Wikipedia

The authors of today's paper exploited the idea in a much wider context. In their previous work this year, **Learning bilingual word embeddings with (almost) no bilingual data**, Artetxe et al. propose to essentially learn every word in a language with this method. In this framework, we never use explicit rules to tell the machine years and months appear around numerals. Instead, we map each word into a continuous high-dimensional space with the word2vec method, so that each word corresponds to a long vector (the authors used 300 dimensions). This representation vector is called an **embedding** because it is learned by inspecting the words around the word of our interest.

After we have determined the embeddings of numerals in our interested languages X and Y, we can stretch and rotate the vectors to make them match. We shall let W denote the linear transformation that maps X to Y, and use the summed least square distance between XW and Y as the loss function. Then, any optimization method from gradient descent to singular value decomposition can be applied to find W. After that, we have two vocabularies with similar embeddings, and we can thus identify the nearest pairs and assign them the same meaning, just like how we assigned the meaning of years and months to the characters around the numerals.

However, since the vocabulary lies in a much larger space, only several seeds cannot help us to match the whole vocabulary accurately at once. The authors propose to only find a few correspondents we are most confident in, and use the larger number of anchors to expand our matching until most of the words are matched. Notably, the process is just like Champollion's progressive deciphering: from the consonants to more pharaohs' names and finally nouns and verbs.



Cross-Lingual Embedding Matching Algorithm

The cross-lingual embedding algorithm described above lays the foundation for our NMT system. The authors have achieved an accuracy of ~40% on English-Italian and English-German pairs, which is much better than the previous work, but still not good enough to directly feed into traditional NMT networks. For the actual translation to happen, we still need more careful designs on the network.

Basic Structure

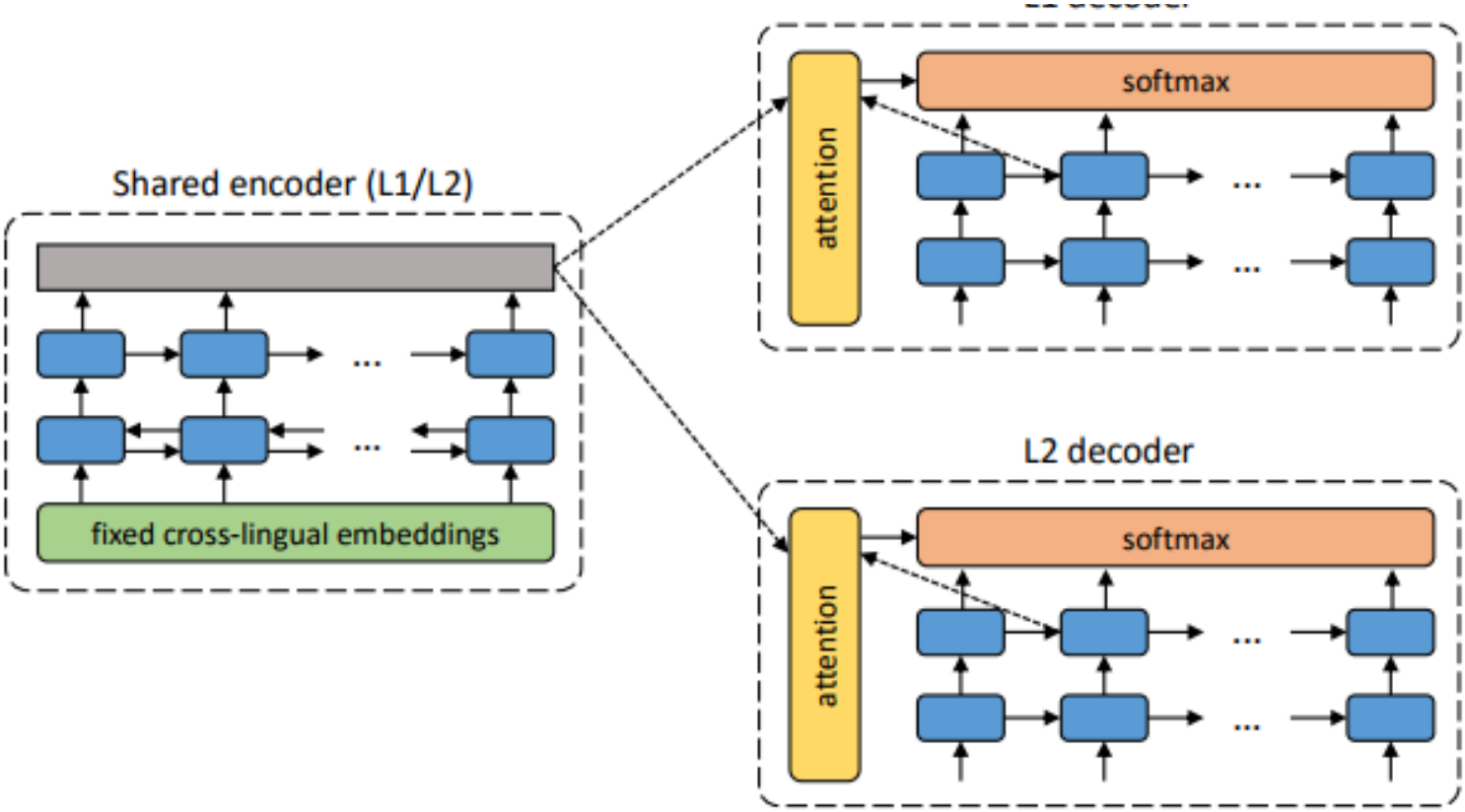
The authors first adopt the state-of-the-art supervised NMT network. The network is composed of three gated recurrent unit (GRU) networks, an alternative version of the LSTM network. One network is the language-to-representation encoder, and another two are the representation-to-language decoders. Each network contains a two-layer GRU with 600 hidden units, and the encoder network also uses a bidirectional structure to better capture the language structure. The global attention method is also introduced to improve the translation quality. With this method, the network only pays the most attention to the relevant phrases in the source text when it generates the translation.

Adapted Features

The authors have adapted some special features to the base network:

- **Dual structure**: the matching between embeddings should be inherent to both languages, so we shall not train the network on one-directional translation only. The authors propose to use the multi-way translation framework by Firat et al.
- **Shared encoder**: both languages share the same encoder to produce the uniform representation of the text regardless of the language. This is possible after we have matched the embeddings in different languages. However, the embeddings for decoding should be different to generate unambiguous texts in different languages.
- **Fixed embeddings** in the encoder: most NMT networks update embeddings along with hidden units during training, but the update will interfere the pre-trained matching, so the embeddings in the encoder are fixed during the training.

With these features at hand, the network looks like the following. Here, softmax refers to the function used to decode, and blue boxes represent hidden units in the network.



Unsupervised NMT Network Architecture

When we train an unsupervised NMT network, we are assumed to have separate monolingual data only, so we do not know what the target translation should be like. The authors thus come up with the following strategies:

- **Denoising**. Without reference data, we can still optimize the system by **self-reconstruction**. Let the shared encoder encode a sentence, and reconstruct the sentence using the decoder for that language. Since we have fixed the embedding in the encoder, it should be able to extract the internal structure from both languages. However, in this setting, the optimal solution would become to copy-and-paste the input, and the resulting translation would be similar to word-to-word substitution. To prevent that degeneration, the authors propose to add random noise like **word swapping** to the input, and use the original sentence as the target output. This way, the encoder can better grasp the information, and become resilient to the change of word orders that are present in many languages. The idea is partly due to the denoising autoencoders by Vincent et al.
- **Backtranslation.** Training through input reconstruction on each separate language does not involve real translation, in which two languages should be closely connected. To fix it, the authors also incorporate the backtranslation training by Sennrich et al: let the raw NMT system to greedily translate from language X to language Y to get some "pseudo-parallel" texts, and then use the other half of the system to backtranslate from Y to X. After that, we can use the similarity between the original sentence and the backtranslated sentence to train the system. Note that we only use real sentences as the benchmark for the output during training, because in translation we only care about whether the output looks like human languages.

As a side note, if we also have access to some parallel data, then we can simply add these to our training, and alternate between the three strategies from batch to batch. As shown in the experiment below, some additional parallel data can greatly improve the translation quality.

With all the techniques introduced above, we can finally do unsupervised translation now!

Experiment Result

Benchmark Comparison

The authors train this network on a standard WMT 2014 task, and then test on the newstest2014 dataset to get the BLEU scores, a commonly used quality metric for machine translation. They tested the system with three settings: unsupervised (as described above; the baseline is the word-for-word substitution using cross-lingual embeddings), semi-supervised (unsupervised + ~100,000 parallel sentence pairs), supervised (traditional method on all parallel data available). The authors also test a popular trick called byte pair encoding (BPE) that shows good result on many traditional NMT networks. The result is given below:

Unsupervised	1. Baseline (emb. nearest neighbor)	9.98	6.25	7.07	4.39
	2. Proposed (denoising)	7.28	5.33	3.64	2.40
	3. Proposed (+ backtranslation)	15.56	15.13	10.21	6.55
	4. Proposed (+ BPE)	15.56	14.36	10.16	6.89
Semi-supervised	5. Proposed (full) + 100k parallel	21.81	21.74	15.24	10.95
Supervised	6. Comparable NMT	20.48	19.89	15.04	11.05
	7. GNMT (Wu et al., 2016)	-	38.95	-	24.61

BLEU performance of each strategy (higher is better)

In summary, the unsupervised translation system works reasonably well, and even beats traditional methods when we have slightly more parallel corpora. The denoising along lowers the performance, but without which we cannot start backtranslation training, while the latter as we see is the key to the performance of the whole system. Meanwhile, the BPE trick contributes little to the performance, and the reason and better solution is yet to be explored.

Sample Translations

As we see below, the network performs well on first two sentences, but makes mistakes on numbers and dates in the third sentence and produce an unreadable translation on the fourth. The mistake on numbers and dates is particularly interesting, because we started our system with cross-lingual embedding of these words. It also makes sense if we think more carefully. With little context at hand, the network can never tell the difference between أكتوبر (Arabic October) and مايو (Arabic May). However, if we have character-level information, we can tell from their corresponding Romanized form--uktūbar and māyū--the system may be better translating them.

Une fusillade a eu lieu à l'aéroport international de Los Angeles.	There was a shooting in Los Angeles International Airport.	A shooting occurred at Los Angeles International Airport.
Cette controverse croissante autour de l'agence a provoqué beaucoup de spéculations selon lesquelles l'incident de ce soir était le résultat d'une cyber-opération ciblée.	Such growing controversy surrounding the agency prompted early speculation that tonight's incident was the result of a targeted cyber operation.	This growing scandal around the agency has caused much speculation about how this incident was the outcome of a targeted cyber operation.
Le nombre total de morts en octobre est le plus élevé depuis avril 2008, quand 1 073 personnes avaient été tuées.	The total number of deaths in October is the highest since April 2008, when 1,073 people were killed.	The total number of deaths in May is the highest since April 2008, when 1 064 people had been killed.
À l'exception de l'opéra, la province reste le parent pauvre de la culture en France.	With the exception of opera, the provinces remain the poor relative of culture in France.	At an exception, opera remains of the state remains the poorest parent culture.

French-English human translation vs machine translation

In the End

I spent half of the time talking about the idea behind deciphering hieroglyphs. I find it exciting to connect historical linguistics to neural networks. Indeed, following the current trend of incorporating everything in AI research, I hope that we can establish more such connections. For example, it is well known in machine translation and foreign language acquisition that English-French is far easier than English-Chinese. However, most work on multi-language machine translation so far does not systematically exploit the linguistic structures shared within groups (e.g. agglutinative languages in the Altaic family, inflectional languages in the Indo-European family). How comparative linguistics could help remains mostly unexplored. If you are interested, feel free to send me an email!

Finally, I would like to thank 十一点半 for his article about ancient character deciphering that inspired this blogpost. He also talks about the deciphering of Linear B and Mayan writing systems. Check it out if you're interested!