Concurrent multilingual Neural Machine Translation of several EU languages

David Dowey *ddowey@stanford.edu*

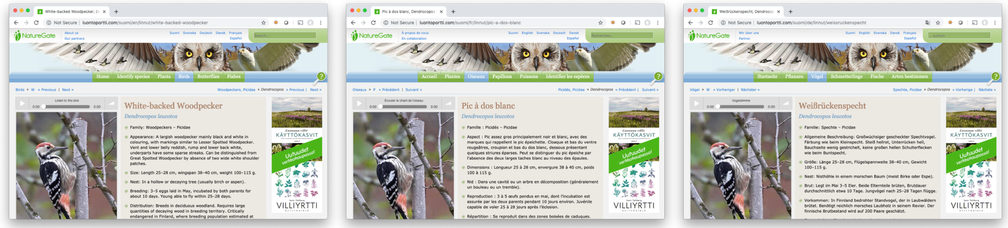
Stanford University

How can we build well-aligned high-quality parallel corpora for NMT in several languages from noisy multilingual websites?

The problem addressed in this project is the messy but useful NLP task of aligning multilingual documents crawled from noisy website sources, for the purposes of generating high-quality parallel corpora of aligned documents across several languages. Automated NMT engines need parallel corpora that are both deep and broad, for example to improve translations tasks in specialized domains, such as for animal and plant species in the example.

The task is not to align the from page content, but rather from the URLs - to use the techniques learned in class to determine which URLs from a scraped multilingual site lead to equivalent content in the different languages. In a sense, we seek to "translate" the URLs, but this is more difficult than it seems, since the URLs contain HTML code, redundant snippets and lots of non-letter characters that don't translate.

English French German



http://www.luontoportti.com/suomi/en/linnut/white-backed-woodpecker

http://www.luontoportti.com/suomi/fr/linnut/pic-a-dos-blanc

http://www.luontoportti.com/suomi/de/linnut/weissruckenspecht

An example is the from the Finnish multilingual nature website luontoportti.com, (Nature Gate).

Site used to identify species of plants, bird, butterflies and fish. Content is in English, French and German, but also Finnish, Swedish, Danish and Spanish.

The site contains lots of media, web elements and links, but the content of interest is the center frame with information on species which is equivalent across multiple languages, which we can use to create domain-specific parallel corpora.

The dataset

- The source data are web-crawled pages along with a language code (EN,FR,DE), the Mime type (text/html), the encoding charset (utf-8), URL, and the HTML and text full page content in Base64 encoding. The examples are scraped, processed and saved as lett files using the the bitextor scraper. The data come from the 2016 WMT shared task on document alignment, extended to multiple languages.
- The input data for the system are extracted from the source data – keeping the text of the webpage URLs and the language code. The URL may be similar across languages, or as in the example, quite different.
- The system outputs the best 1:1:1 match of URLs in EN,FR,DE as a triplet in that order. The URLs can then be used to extract the full page content from the source data to construct the parallel corpora.
- The testing triplets are hand chosen true matches (the oracle) used to measure the accuracy of the system

Domain	English pages	French pages	German pages	Testing triplets
luonoportti.com	3,645	1,796	2,146	70
vinci.com	3,564	3,374	3,005	50
dakar.com	17,420	14,582	1,655	50
krn.org	115	115	115	50
prohelvetia.ch	5,209	4,421	6,289	50
pawpeds.com	1,011	136	86	50
inst.at	3,203	543	17,076	50
schakportalen.nu	33	29	29	23
rehazenter.lu	201	317	161	50
cyberspaceministry.org	1,534	958	173	40
galacticchannelings.com	4,231	1,283	983	50
	40,166	27,554	31,718	533

Processing pipeline

Extract most important terms from multilingual URLs
UCS segmentation algorithm using character bigrams

Output: lang code, URL, important terms
en; URL; ['en', 'linnut', 'white' 'backed', 'woodpecker ']
fr; URL; ['fr', 'linnut', 'pic' 'a', 'dos ', 'blanc']
de; URL; ['de', 'linnut', 'weissruckenspecht ']

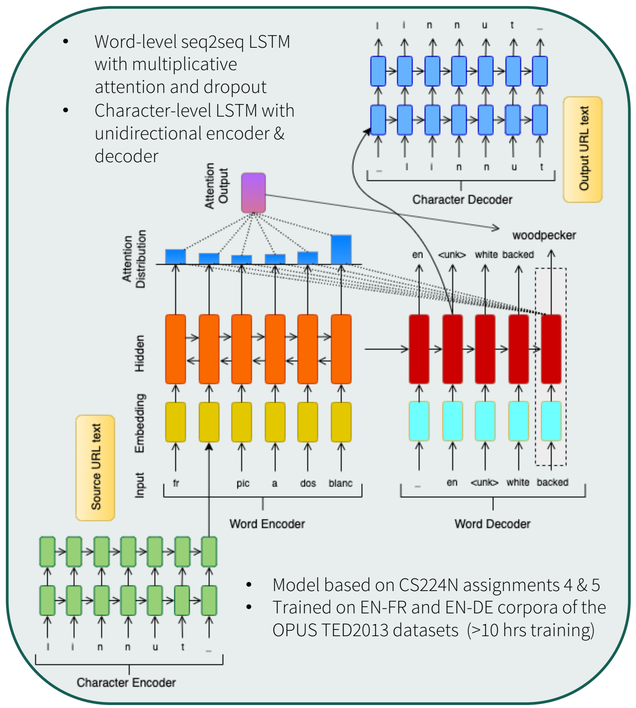
1. We use a modification of the **UCS algorithm from text reconstruction HW3** to disgard common non-informative terms from the URLs and keep the rarer important ones.
2. Build a language model for **character bigram scores** - the first character paired with the next 1 to 6 characters: B(1,1) to B(1,6). This is trained on the input data of multilingual URLs.
3. Segment the text using **UCS and additive bigram scores** – for common terms, B(1,1) is small and these segment to individual letters.

e.g. **h t t p : / / w w w . l u o n t o p o r t t i . c o m / s u o mi/en/lin nut/white -backed-w oodpecker**

4. Throw away single characters, collate the remaining terms and then tokenize using punctuation separators (e.g. '/', '_' or '-')

Translate important terms to their English equivalents
Hybrid character-word seq2seq LSTM Neural Machine Translation model with attention

Output: lang code, URL, translated important terms
fr; URL; ['en', 'linnut', 'white' 'backed', 'woodpecker ']
de; URL; ['en', 'linnut', 'white' 'backed', 'woodpecker']



Match translated target URLs 1:1:1 to their source English counterpart
Search by n-gram precision

Output: the aligned URL triplet
(http://www.luontoportti.com/suomi/en/linnut/white-backed-woodpecker,
http://www .luontoportti.com/suomi/fr/linnut/pic-a-dos-blanc,
http://www.luontoportti.com/suomi/de/linnut/weissruckenspecht)

1. We want to find the best match of the translations, so in the spirit of the **BLEU score**, we search for the triplet which yields the **highest n-gram precision**.
2. Since the URL translations are short, N=3 suffices
3. Start with the language with the fewest pages, for example, the French translation ouput:

e.g. **['en', 'linnut', 'white' 'backed', 'woodpecker ']**

4. Find the best match in English and German and take the original URLS to make a 1:1:1 triplet.

Extract content from URLS

Why use this method?

- Complexity: avoids parsing and translation of the full-text of almost 60,000 webpages, so much faster.
- Steps 1 and 3 are simple and require no training, while the NMT model in Step 2 can be trained on other data.
- To exploit techniques learned in CS221 and CS224N.

What next?

- Add Spanish URLs
- Alternative parallel corpora for pretraining the NMT model
- Extend dataset to more multilingual domains & more testing triplets
- Multilingual BERTor ELMo word embeddings in Step 2.

How well does it work?

- The system performance was measured against the recall score of **the baseline method** using the **minimum edit distance** of the URLs, thresholded at a maximum of 10 character edits, to avoid repetitive matches.
- The system achieves **450 out of 533** possible unique 1:1:1 testing triplets, achieving **84.4% recall**, compared to a baseline score of 354 and 66.4% recall.
- **Error-analysis** shows system performs poorly when English URLs contain colloquial terms (e.g. for birds, fish) and if many English pages are duplicates.

References

Christian Buck and Philipp Koehn. Findings of the WMT 2016 bilingual document alignment shared task. In Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers, pages 554–563, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-2347. URL https://www.aclweb.org/anthology/W16-2347.

Jakob Uszkoreit, Jay Ponte, Ashok Popat, and Moshe Dubiner. Large scale parallel document mining for machine translation. In Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 1101–1109, Beijing, China, August 2010. Coling 2010 Organizing Committee. URL https://www.aclweb.org/anthology/C10-1124.

J. Tiedemann, 2012, Parallel Data, Tools and Interfaces in OPUS. In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)