Abstract

Source language detection

mention that it isn't a common task; related work here too

motivation: test sota models, if detection possible, we're curious which features will be salient (xd) for detection

Gather lots of text documents in a few languages, use machine translation models to translate them to english, feed them to some kind of model, telling it what language each document was translated from. See if it can learn to recognize that. If the translation models are good enough, that should be impossible.

some previous papers have shown it's possible to detect the original language of a human-generated translation

Related work

https://aclanthology.org/2021.naacl-main.462.pdf It detects the translated text using round-trip method.

https://arxiv.org/pdf/1910.06558.pdf . It uses back translation method.

https://aclanthology.org/W18-1603.pdf $^{t}ranslatedtextdetectiononChinese.$

https://www.cs.cmu.edu/ dkurokaw/publications/MTS-2009-Kurokawa.pdf

$^{T}hispaperdetectstexttranslatedfromfrench, theysaysomen-gramswereveryfrequent, andalsomorearticlesandprepositionsthaninoriginallyenglishtext$

"good classification accuracy was obtained even when texts were reduced to part-of-speech sequences" maybe use some model based on POS sequences, then?

https://aclanthology.org/C12-2076.pdf

$^{h}ispaperisinterestingbecausethetaskissimilartoours.Amongotherstheycreatevectorrepresentationsforarticlesbasedonse$ $levelmetrics, andSVMbasedonthat.Certain2-gramswereveryfrequentfortranslationsfromcertainlanguages$

Maybe easier to recognize longer text (for reliable document-level statistics), which is why we use whole paragraphs rather than sentences

Approach

relevant to the lecture becaaause 1. we use deep learning 2. we evaluate sota deep learning

chosen languages grammar not similar to english configurational languages?

dataset creation method, applied models

some paragraphs shorter because removed sentences of length >256 after tokenization random link sampling + at most two paragraphs from each site to avoid too many related to the same subject decided not to remove proper names even though one paper did. Just limited the number of paragraphs from the same site; there was really lots of diversity, and besides there was an overlap in subjects between languages (e.g. those pesky christians in both arabic and indonesian datasets) so we decided it's safer to just leave them, especially since otherwise we'd have had to replace them with something so that all the grammar of the sentence doesn't go bonkers (especially after translation)

paragraphs are not actually that - all sentences in a given article are concatenated together, and then we create two chunks by choosing two sequences of whole consecutive sentences, so that the length of a chunk (in words) only slightly exceeds 256 (i.e. would be below 256 if we didn't include the last sentence).

indonesian dataset: 252 from deepl 995 from microsoft 330 from libretranslate $^{n}umbersbeforeremovingduplicates, inthewholeIndonesianset, there$

dataset: mention the proper names discussion

for trees: https://aclanthology.org/P15-2029.pdf $^{d}ependencytreeCNN(?)concatenatingancestralvectors(finalmethod$

https://nlp.stanford.edu/pubs/zhang2018graph.pdf $^{a}lternativemethod$

https://arxiv.org/pdf/1609.03286.pdf $^{a}lsoprocessingparsetrees$

and explain who chose one-hot POS embedding and not to include siblings oh and why dependency parsing rather than abstract meaning representation (why? syntax)

Results

introduce the test results, draw some conclusions

Conclusion

sum up, propose further work, acknowledge shortcomings

Work distribution

- Chih-Hsiang Hsu
  - todo
  - todo
- Chung-Hao Liao
  - todo
- Antoni Maciąg
  - todo
- Jen-Tse Wei
  - todo