Abstract

Source language detection

Although there is some preexisting work on tasks similar to the one we chose for the assignment, it is not a standard NLP task and to the best of our knowledge, there is no work with exactly the same problem formulation. The formulation is: given a text machine-translated into English from a known set of source languages, detect the source language. We will refer to the problem as Source Language Detection (SLD).

We are motivated by the following: in human translation, clues as to the original language in the form of both syntactic and semantic information tend to get unconsciously carried over to the translated text [1], making it possible to detect the original language. We are curious if that is also the case for currect state-of-the-art models for Machine Translation, and if so, what kinds of models will make SLD possible and what features of the translated text will be salient for detection. If the translation models are good enough, our obtained accuracy should not be considerably higher than random guessing. In addition, we are curious if there is a difference between translation models which were trained multilingually and ones that were not. We hypothesize that for the former, the task might be more difficult because such models have learned on languages with various syntactic structures, which could make them less likely to carry the syntactic features of a particular source language over into the translation.

Related work

Nguyen-Son et al. [2] detect the correct one out of a set of possible source language - translator tuples, with possible languages being Russian, German and Japanese, using the round-translation method. It utilizes the phenomenon by which, while repeatedly translating a text back and forth between two languages, each round-trip changes the text less than the previous one. Thus, given an English text T which we know was translated from either Russian or German, if we generate round-trip $En \to Ru \to En$ and $En \to Ge \to En$ translations of the text, the similarity to T will be higher for the translation through the language that was the original language of T.

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.

we have not been able to find an attempt to create a source language detector oblivious of the used translator - no wait, the round-trip does that

in comparison to Son, ours would be much faster if it works and oblivious to the translator used, both in training (multiple subclassifiers) and in recognition (multiple translations)

https://aclanthology.org/W18-1603.pdf

$^t ranslatedtextdetectiononChinese, dependencytrees$

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

$^T hispaperdetectstexttranslatedfromfrench, theysaysomen-gramswereveryfrequent, andalsomorearticlesandprepositionsthani$

"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

$^h ispaperisinterestingbecausethetaskissimilartoours.Amongother$ $levelmetrics, andSVMbasedonthat.Certain2 - gramswereveryfrequentfortranslationsfromcertainlanguages$

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

Approach

Chosen languages

chosen languages grammar not similar to english configurational languages?

Dataset creation

a comparison between multilingually trained models and not 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), and besides for POS-based models, that doesn't make a difference either way

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).

Final dataset composition for each language (everone should describe their own, if possible):

- Arabic:

- Chinese:

- Indonesian: 252 from deepl 995 from microsoft 330 from libretranslate $^{n}umbersbeforeremovingduplicates, inthewholeIndonesianset, therewereabout2kJapanese$

Models

Four models have been created: blah blah lenin was a mushroom

more precise descriptions (everyone should describe their own):

Bert

RoBERTA on POS tags

SVM

Dependency tree CNN

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) explain how sentence length, number of ancestors was chosen

Results

introduce the test results, draw some conclusions for every language-translator pair, number of correctly/incorrectly classified paragraphs, if possible. also ofc everyone should report their own

Conclusion

sum up, propose further work, acknowledge shortcomings

If it turns out our model is trash, it might be either because 1. It really is trash 2. current state-of-the-art translation models are just so good

maybe we should try it on some old translation model, worse than current SOTA

for validation we maybe should have different translators, so that we're sure our model learned what text translated from Korean looks like, and didn't just learn what text translated by Google Translate looks like. we ended up not doing that. But we did include other models for the train set to make it noisier. So maybe it's not that bad.

Work distribution

I really hope I didn't make any mistakes in your names XD

- Chih-Hsiang Hsu
  - todo
  - todo
- Chung-Hao Liao
  - todo
- Antoni Maciąg
  - todo
- Jen-Tse Wei
  - todo
- Each member:
  - Writing the part of the report about their respective dataset part and model.

References

[1] Gerard Lynch and Carl Vogel. Towards the Automatic Detection of the Source Language of a Literary Translation. Proceedings of COLING 2012

[2] Hoang-Quoc Nguyen-Son, Tran Phuong Thao, Seira Hidano, Ishita Gupta, and Shinsaku Kiyomoto. Machine Translated Text Detection Through Text Similarity with Round-Trip Translation Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics, 2021

[3] Hoang-Quoc Nguyen-Son, Tran Phuong Thao, Seira Hidano, and Shinsaku Kiyomoto. Detecting Machine-Translated Text using Back Translation. Proceedings of the 12th International Conference on Natural Language Generation, 2019

[4] Hai Hu, Wen Li, Sandra Kubler. Detecting Syntactic Features of Translated Chinese. Proceedings of the Second Workshop on Stylistic Variation, 2018

[5] David Kurokawa, Cyril Goutte and Pierre Isabelle. Automatic Detection of Translated Text and its Impact on Machine Translation. In Proceedings of Machine Translation Summit XII, 2012

[6] Mingbo Ma, Liang Huang, Bowen Zhou, Bing Xiang. Dependency-based Convolutional Neural Networks for Sentence Embedding. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 2015

[7] Yuhao Zhang, Peng Qi, Christopher D. Manning. Graph Convolution over Pruned Dependency Trees Improves Relation Extraction. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018

[8] Yun-Nung Chen, Dilek Hakkani-Tur, Gokan Tur, Asli Celikyilmaz, Jianfeng Gao, and Li Deng. Knowledge as a Teacher: Knowledge-Guided Structural Attention Networks, 2016