Suid-Afrikaanse Tydskrif vir Natuurwetenskap en Tegnologie
ISSN: (Online) 2222-4173, (Print) 0254-3486

Evaluasie van skoner vertaalgeheues

Outeurs:
Friedel Wolff,
L Pretorius,
P Buitelaar

Affiliasie:
Skool vir Rekenaarkunde,
Universiteit van Suid-Afrika

Korresponderende outeur:
F Wolff
E-pos:
friedel@translate.org.za
Smutslaan 172, Lyttelton Manor, 0157

Hoe om hierdie artikel aan te haal:
Friedel Wolff, L Pretorius, P Buitelaar, Evaluasie van skoner vertaalgeheues, *Suid-Afrikaanse Tydskrif vir Natuurwetenskap en Tegnologie* 38(1) (2019). https://doi.org/10.36303/SATNT.2019.38.1.782

Evaluation of cleaner translation memories: We adapt a supervised learning method to operate as an unsupervised method identifying erroneous entries in translation memories. Artificial training data is generated to make this possible. We show improved results in a translation memory system and neural machine translation systems with the datasets cleaned by removing the erroneous entries.

Rekenaars kan 'n nuttige rol speel in vertaling. Twee benaderings is vertaalgeheuestelsels en masjienvertaalstelsels. By hierdie twee tegnologieë word 'n vertaalgeheue gebruik – 'n tweetalige versameling vorige vertalings. 'n Vertaalgeheuestelsel herroep vorige vertaalde segmente wat 'n vertaler kan bystaan in die huidige teks wat vertaal word. By masjienvertaling word 'n teks outomaties van die brontaal na die doeltaal vertaal. Hierdie navorsing doen ondersoek na hoe so 'n vertaalgeheue se kwaliteit verbeter kan word, en bepaal of 'n skoner vertaalgeheue meer geskik is vir 'n spesifieke taak as vantevore.

Deur middel van masjienleer op 'n geannoteerde datastel kan 'n klassifiseerder gebou word om die kwaliteit van 'n vertaalgeheue te verbeter deur foutiewe inskrywings in 'n vertaalgeheue te identifiseer (Wolff 2017). Deur die foutiewe inskrywings uit 'n vertaalgeheue te verwyder, is 'n kleiner, skoner vertaalgeheue beskikbaar vir toepassings. Weens die gebrek aan gepaste data vir opleiding, asook die ongebalanseerde klasverdeling (die verhouding korrek:foutief) in tipiese datastelle, blyk realistiese gebruik in bogenoemde toepassings nie triviaal te wees nie.

Die gekontroleerde benadering word uitgebrei om as 'n ongekontroleerde benadering te funksioneer. Tegnieke vir gekontroleerde leer soos steunvektor-, ewekansigewoud- en logistieseregressie-klassifiseerders word steeds gebruik, maar kunsmatige opleidingsdata word gegenereer waarmee die gewenste skaal bereik word en die klasverdeling na willekeur bepaal kan word. Hiermee kan 'n parameter dus beïnvloed hoe streng die resulterende klassifiseerder is. Die kunsmatige opleidingsdata word gegenereer deur die doelteks van ongeannoteerde vertalings te versteur deur 'n kunsmatige fout in die vertaling te genereer. Voorbeelde van sulke foute is die byvoeging of verwydering van 'n letter, 'n woord, 'n leesteken, of selfs 'n groter deel van die segment. Die datastel wat aangevul word met die kunsmatige foutiewe vertalings dien dan as opleidingsdata vir die leermetode.

Die navorsing word intrinsiek en ekstrinsiek geëvalueer. Die intrinsieke evaluasie dui op 'n stelsel wat die klassifikasieakkuraatheid van die gekontroleerde benadering grotendeels benader. Vir ekstrinsieke evaluasie dui eksperimente aan dat so 'n vertaalgeheue beter prestasie in 'n vertaalgeheuestelsel behaal as die vuil basislyn. Neurale masjienvertaalstelsels is bekend daarvoor dat hulle sensitief is vir die kwaliteit van die opleidingsdata, en dat reusehoeveelhede data nodig is om 'n kompeterende stelsel op te lei. As ondersteunende bewys vir die waarde van 'n skoner vertaalgeheue word 'n verbetering ook aangedui by die opleiding van 'n neurale masjienvertaalstelsel, ten spyte van minder opleidingsdata in die skoner datastel.

Literatuurverwysings

Wolff, F., 2017, 'Combining off-the-shelf components to clean a translation memory' *Machine Translation,* 30, 167–181.

Nota: 'n Seleksie van referaatopsommings: Studentesimposium in die Natuurwetenskappe, 25–26 Oktober 2018, SA Akademiegebou, Pretoria, Suid-Afrika. Gasredakteurs: Prof Rudi Pretorius (Departement Geografie, Universiteit van Suid-Afrika); Prof Chris Swanepoel (Departement Besluitkunde, Universiteit van Suid-Afrika); Me Andrea Lombard (Departement Geografie, Universiteit van Suid-Afrika)