What's in a Domain? Towards Fine-Grained Adaptation for Machine Translation.

M.E. van der Wees

Samenvatting

Computervertaalsystemen (Engels: machine translation (MT) systems) gebruiken software om teksten geschreven in één taal (bijvoorbeeld Nederlands) te vertalen naar een andere taal (bijvoorbeeld Engels). Hedendaagse computervertaalsystemen worden gebouwd door te leren van grote hoeveelheden voorbeeldvertalingen tussen twee talen, zogeheten *parallelle corpora*. Als zulke parallelle corpora voldoende groot zijn en van goede kwaliteit, vaak vertaald door professionele vertalers, kan een computervertaalsysteem vertalingen van goede kwaliteit genereren.

Parallelle corpora van goede kwaliteit zijn echter enkel beschikbaar voor een beperkt aantal *domeinen*, zoals nieuws of parlementaire verslaggevingen. De situatie is minder rooskleurig voor domeinen waarvoor meertalige *training data* schaars is, wat bijvoorbeeld het geval is voor medische teksten of berichten op sociale media. In zulke gevallen zorgen de verschillen in schrijfstijl en vocabulaire tussen de voorbeeldvertalingen en de te vertalen tekst ervoor dat de vertaalkwaliteit drastisch verslechtert. In de afgelopen jaren is er veel aandacht besteed om dit probleem op te lossen door middel van domeinadaptatie (Engels: domain adaptation). In dit proces wordt het vertaalsysteem aangepast aan het domein van belang waardoor de vertaalkwaliteit voor dit domein verbetert.

Helaas is het concept *domein* niet eenduidig gedefinieerd in de huidige literatuur. Meestal betekent een nieuw domein een 'andere dataset' en wordt informatie over de oorsprong van deze dataset direct gebruikt om een vertaalsysteem te optimaliseren. Deze definitie negeert drie belangrijke feiten: ten eerste, documenten of zinnen binnen een domein kunnen variëren op vele verschillende niveaus, zoals qua *onderwerp*, *tekst-genre* of *taalgebruik*. Variatie op deze niveaus kan belangrijke informatie bevatten om een vertaalsysteem aan te passen. Ten tweede, sommige domeinen of tekst-genres vereisen specifieke strategieën om vertaalkwaliteit te verbeteren dankzij inherente kenmerken van die domeinen. Ten derde, beschikbare domein-labels bevatten niet per definitie de meest waardevolle informatie voor effectieve adaptatie.

Om inzicht te krijgen in het concept domein en de impact van domeinen op computervertaalsystemen, stelt dit proefschrift de vraag: "Wat omvat een domein?" Aan de hand van deze vraag onderscheiden we verscheidene aspecten die tezamen een domein definiëren, zoals onderwerp, tekst-genre, taalgebruik, dialoogkenmerken, sprekers en het geslacht van sprekers. We bestuderen in hoeverre vertaalkwaliteit varieert binnen elk van deze aspecten, en hoe we deze aspecten kunnen gebruiken om adaptatie van vertaalsystemen op verschillende niveaus te bewerkstelligen. Hierbij zijn we specifiek geïnteresseerd in *informele* teksten en *conversaties*, die beide worden gekenmerkt door een gebrek aan standaardisatie en een notoir slechte vertaalkwaliteit. Daarnaast beogen we methodes te ontwikkelen die niet, of slechts gedeeltelijk, afhankelijk zijn van handmatig gedefiniëerde domein-informatie.

Door te bestuderen wat een domein omvat en aan te tonen hoe we verschillende aspecten van taal kunnen benutten om computervertaalsystemen te verbeteren, nemen we in dit proefschrift een belangrijke stap richting verbeterde adaptatie voor computervertaalsystemen.