Sprogteknologi er overalt og forandrer langsomt måden, vi arbejder på, måden, vi kommunikerer med hinanden, måden, vi lagrer viden. Sprogteknologi er søgemaskiner, Siri, Google Now. Det er ordbøger i mobiltelefoner, maskinoversættelse, de tale-til-tekst-systemer, som læger bruger til at skrive notater i patientjournaler, og det er stemmen i metroen eller bilens GPS.

SPROGTEKNOLOGI FOR RESTEN AF VERDEN

ANDERS SØGAARD

CENTER FOR SPROGTEKNOLOGI, KØBENHAVNS UNIVERSITET

Det er understøttende software i sprogundervisning, teknologier til at hjælpe ordblinde, det er værktøjer, som professionelle oversættere bruger til at effektivisere deres arbejdsgang. Det er programmer, som virksomheder bruger til at finde ud, hvad deres kunder siger om deres produkter på de sociale medier. Det er programmer, der automatisk genererer billedtekster. Software til automatisk resumering af store tekstsamlinger. Markedet for sprogteknologi – virksomheder, der udvikler den her slags produkter – vokser og vokser. Det skyldes både store virksomheder som Google og Yahoo, men også hundreder og tusinder af mindre start-ups.

Problemet er, at sprogteknologi kun udvikles for en brøkdel af verdens sprog. I dag findes langt de fleste sprogteknologier kun for 20-30 sprog. Der er omkring 7000 sprog i verden. En del af disse sprog er mindre, og flere tusind af dem har ikke et skriftsprog. Men iblandt de sprog, vi ikke har sprogteknologi for, er også en lang række store sprog, som er millioner af menneskers modersmål, og ofte eneste sprog, f.eks. telugu (75 millioner), marathi (73 millioner) og hausa (50 millioner). I Indien er det kun omkring 6%, der taler engelsk. Udover engelsk har vi sprogteknologi for hindi, som tales af omkring 300 millioner indere. Resten af inderne er praktisk talt uden sprogteknologi, og de

fordeler sig på omkring 1650 sprog og dialekter. Selv i Europa er der store sprog uden sprogteknologi, som f.eks. armensk (8 millioner) og albansk (6 millioner).

Jeg har – med en lang række af kollegaer fra Google, MIT, Columbia, Edinburgh – længe arbejdet med at udvikle sprogteknologi for flere sprog, under overskriften cross-lingual learning. Sprogteknologi er anvendt af maskinlæring – og altså funktionsestimering fra observationer. Observationerne er i det her tilfælde lingvisters manuelle analyser af sætninger og tekster. Et program, der kan bestemme ords syntaktiske kategori, er en funktion fra sekvenser af ord til sekvenser af kategorier – og for at kunne bestemme en sådan funktion skal man have et sample af sætninger annoteret med analyser. Og det er netop det, der er flaskehalsen.

I cross-lingual learning forsøger vi at bruge manuelt analyserede tekster på ét sprog til at estimere funktioner for et andet sprog. Vi estimerer altså funktioner ud fra biased data – data, der ikke bare er skævt samplet, men samplet fra en anden, men relateret distribution. Indtil for nylig eksperimenterede vi med kendte sprog, men lod, som om vi ikke havde data for ét af dem. Altså, vi brugte manuelt analyserede tekster fra engelsk og tysk til at lære modeller for spansk. Vi kunne så bruge manuelt analyserede tekster fra spansk (som vi i læringsfasen lod som om ikke fandtes) til at evaluere vores modeller.

Der er flere problemer med at lave cross-lingual learning med de 20-30 sprog, for hvilke vi har manuelt analyserede tekster. For det første er de alle sammen tæt beslægtede. De 20-30 sprog, vi har manuelt analyserede tekster for, er næsten alle indo-europæiske sprog og dermed tættere beslægtede end sprogpar i almindelighed. De indo-europæiske sprog er også lettere at arbejde med end andre sprog, fordi de bruger mellemrum til at afgrænse ord. Det er altså lettere for vores modeller at finde analysens mindstedele. På mange andre sprog bruger man ikke mellemrum til at afgrænse ord. Endelig har en del arbejde indenfor cross-lingual learning brugt oversat tekst til at lære relationen mellem forskellige sprog. Og her er der endnu et bias. Der findes langt flere oversættelser mellem de indo-europæiske sprog, end der findes for sprogpar på tværs af sprogfamilier.

En del af vores forskning lige nu handler om rent faktisk at udvikle sprogteknologi for alle verdens sprog – og ikke blot udvikle sprogteknologi for de 20-30 sprog uden at bruge de manuelt analyserede tekster, vi har til rådighed. Hvad findes der af ressourcer for disse sprog? Det varierer meget. For nogle af sprogene findes der ordbøger, wiki'er, oversættelser af resolutioner, literære værker og localization-filer, og referencegrammatikker. For andre findes der ikke meget andet end en oversættelse af det Ny Testamente, et par numre af Vagttårnet og en masse hjemmeside og tekster fra sociale medier.

"Problemet er, at sprogteknologi kun udvikles for en brøkdel af verdens sprog."

I vores første eksperimenter har vi kun antaget, at der findes en oversættelse af det Ny Testamente – en antagelse, der gælder for omkring to tusinde sprog. Vi har udviklet en teknologi, hvor vi først lærer sprogteknologiske modeller for de sprog, for hvilke vi har manuelt analyserede tekster, dernæst analyserer det Ny Testamente, dernæst projicerer vores viden fra et sprog til et andet, så vi kan bestemme sandsynlighed for ethvert ords syntaktiske kategori – også for de sprog, vi ikke har manuelt analyserede tekster for. På baggrund af annoteringen af det Ny Testamente kan vi så estimere funktioner for alle to tusinde sprog. Når det gælder bestemmelse af ords syntaktiske kategori, kan vi med denne teknologi estimere funktioner med omkring 80% nøjagtighed – altså, som tildeler fire ud af fem ord den rigtige syntaktiske kategori. Hvis vi antager lidt mere data – f.eks. oversættelser af FN-resolutioner, et par numre af Vagttårnet, eller lignende – får vi bedre resultater. Hvis vi antager eksistensen af en ordbog, får vi endnu bedre resultater.

Visionen er at udvikle sprogteknologi for alle verdens sprog. Problemet er et af de sværeste indenfor anvendt maskinlæring lige nu, og motivationen er dels de tekniske og metodologiske udfordringer, men også at sprogteknologi lige nu giver en lille del af verden et kæmpe forspring – og måske ikke den del af verden, der mest af alt har brug for det forspring.