Språkbanken, Nasjonalbiblioteket, 03.12.2020

Målfrid: Måling av bokmål og nynorsk på statlige nettsider

Dette notatet presenterer metode, funn, metodiske forbehold og planer for videre arbeid basert på en forstudie i Målfrid-prosjektet der Språkbanken høster og prosesserer materiale fra statlige nettsider for å tallfeste fordelingen av bokmål og nynorsk. Målingen skal foregå minst to ganger i året og fremskaffe grunnlagsdata for Språkrådets rapporteringsarbeid i forbindelse med Mållova på en mer effektiv måte enn tidligere.

1. Metode

Målfrid inkluderer høsting av nettmateriale, prosessering og språkdeteksjon. I det følgende beskrives de forskjellige stegene.

a) Høsting

Første steg i Målfrid-prosessen er selve nedlastingen eller høstingen av statlige nettsider, som skal være så omfattende som mulig. Høstingen er gjort med det fritt tilgjengelige og godt konfigurerbare verktøyet *wget* (https://www.gnu.org/software/wget/). Målfrid-høsteren er konfigurert til å gå gjennom og laste ned alle sider innenfor et gitt domene rekursivt ned til nivå 12 (f.eks. vil den i tilfelle UiO inkludere såkalte subdomener som hf.uio.no, men alltid holde seg innenfor domenet uio.no og ikke ta med sider som er lenket til utenfor uio-domenet). Vi høster ned til nivå 12, som var det dypeste nivået vi fant i en teststudie av utvalgte store domener. Vi har satt en absolutt grense for å unngå uendelige løkker, noe som er potensielt problem i forbindelse med rekursiv nedlasting av nettsider. Fra Brønnøysundregistrene har vi fått overlevert lister med domener tilknyttet statlige institusjoner. Høstingen utelukker visse metadataregistre og databaser som kun inneholder katalogdata og ikke naturlig språk i form av setninger. Domenene og underdomenene som er utelukket, befinner seg i block_list.xlsx. Denne listen er utarbeidet av Språkrådet, etter en kvalitativ gjennomgang av data fra Nasjonalbiblioteket. Å utelukke domener som ikke omfattes av Mållova, har bidratt til en betydelig forkorting av innhøstingstiden.

b) Prosessering

Etter selve høstingen sitter vi igjen med en stor mengde filer. For materialet i denne pilotstudien dreier det seg om ca. 13,5 millioner dokumenter med tekstinnhold (f.eks. .html = 12,9 millioner, .pdf = 426000 og .doc(x) = 21.000). I pilotstudien ser vi på HTML-dokumenter, altså nettsider, som utgjør brorparten av materialet. Prosesseringen består av følgende steg:

- HTML-prosessering: “Boilerplate removal”

Den største gruppen av dokumenter på nettet, HTML-dokumentene, har mye innhold som ikke er naturlig språk (f.eks. elementer og lenker). I tillegg er det en del naturlig språk som repeteres i form av menytekst, rammetekst, topp- og bunntekst osv. For å bøte på dette, brukes en metode kalt “boilerplate removal” hvor kun innhold i form av overskrifter og hele avsnitt hentes ut. I Målfrid benyttes algoritmen jusText (http://corpus.tools/wiki/Justext) som bruker flere heuristikker for å finne "godt" innhold: bla. semantikken i HTML-elementer (et <nav>-element tyder på menyvalg osv.), tettheten av såkalte stoppord (høyfrekvente ord som *og, jeg, er, var, som* osv.) som gir en pekepinn på om hele setninger forekommer, og tettheten av lenker. Etter “boilerplate removal” sitter vi igjen med ren tekst, fordelt på avsnitt.

- Deduplisering

Når teksten er hentet ut av HTMLene og PDFene, får denne teksten en såkalt sjekksum (her brukes hashing-algoritmen sha256). På denne måten kan vi identifisere identiske dokumenter innenfor enkelte institusjoner eller i hele tekstmengden. Identiske dokumenter, som det er mange av, teller kun én gang innenfor en statlig institusjon. Det kan være flere grunner til at ett og samme dokument forekommer flere ganger, men det har ofte å gjøre med hvordan nettsiden er organisert.

- Tekstekstraksjon fra andre dokumenter

I den endelige leveransen vil tekst også ekstraheres fra PDF- og Worddokumenter, hvor særlig førstnevnte krever spesialbehandling. PDF-filer er digitalt fødte eller skannede dokumenter. PDF er et papirorientert format og ikke særlig egnet til maskinell prosessering av språkdata fordi det er så tett bundet til en fysisk side. Det er mulig å hente ut tekst fra PDFer som inneholder et såkalt tekstlag, men dette tekstlaget er ofte av svært varierende kvalitet og i tekstdokumenter med flere spalter (f.eks. Stortingsdokumenter) går spaltene over i hverandre på linjenivå. Derfor har vi besluttet at alle PDF-dokumentene i Målfrid skal OCR-behandles. OCR-behandlingen av alle PDFer fra første innhøstingsrunde (ca. 425.000 PDFer) er estimert til å koste ca. 40.000 kroner på Google Cloud Platform og behandlingen vil ta ca. en uke. Suksessive innhøstinger vil ha et betydelig lavere antall PDFer siden svært mange vil være duplikater fra første innhøsting.

c) Språkdeteksjon

Til slutt kjøres språkdeteksjon på de unike dokumentene. Her brukes algoritmen TextCat (https://www.let.rug.nl/vannoord/TextCat/) i implementasjonen Pytextcat fra Giellatekno og med språkmodeller fra samme sted (https://giellalt.uit.no/ling/langrec.html). TextCat ser på forekomsten av såkalte bokstav-ngrammer og hele ord opp mot en modell og bruker en såkalt “out-of-place rank” for å sammenlikne et dokument med en modell. Dette er et enkelt mål som er svært etablert innenfor språkdeteksjon og er beskrevet i en artikkelen av Cavnar og Trenkle (1994), *N-Gram-Based Text Categorization* (https://www.let.rug.nl/vannoord/TextCat/textcat.pdf).

Språkdeteksjonen kjøres på to nivå: på dokumentnivå og på avsnittsnivå. I målingen på dokumentnivå får hvert dokument den mest sannsynlige språkklassifiseringen etter modellen

over. I målingen på avsnittsnivå får hvert avsnitt en egen klassifisering og slik kan man ta hensyn til flerspråklighet i dokumenter. Kun avsnitt med en lengde på mer enn 25 ord tas med i beregningen fordi språkdeteksjon på korte sekvenser er problematisk (se diskusjon). For mange / de fleste dokumenter i korpuset vil nok språkdeteksjon på dokumentnivå være tilstrekkelig, men på nettsider med flerspråklige dokumenter og i andre dokumenter (f.eks. visse Stortingsmeldinger) vil det være hensiktsmessig med språkdeteksjon på avsnittsnivå.

Til sist har vi valgt å bruke Språkrådets metode for tidligere rapportering på det samme materialet for å ha et bedre sammenligningsgrunnlag. Vi hentet ut dokumentfrekvens for henholdvis *ikke* og *ikkje* og *fra* og *frå* og beregnet nynorskandelen på basis av disse (ref. https://www.sprakradet.no/Spraklige-rettigheter/Spraklege-rettar-som-gjeld-bruken-av-norsk/Maallova/Rapport/).

2. Funn

Funnene av pilotstudien presenteres i form av regneark. Regnearkene viser andelen av nynorsk i alle HTML-dokumenter som er klassifisert som norske i materialet som er lastet ned fra de statlige institusjonene. Dokumenter på andre språk er ført opp under «other» og er ikke tatt med i beregningen av nynorsk- og bokmålsandel. I forbindelse med pilotstudien gis resultater fra de tre metodene beskrevet ovenfor:

- all-institutions-textcat-docs.xlsx - tekst klassifisert på dokumentnivå, viser antallet dokumenter på bokmål, nynorsk og andre språk (metode: Textcat)
- all-institutions-textcat-paragraph-tokens.xlsx - tekst klassifisert på avsnittsnivå, viser antallet ord som er klassifisert som bokmål, nynorsk og annet (metode: Textcat)
- all-institutions-wordmethod-docs.xls - viser antallet dokumenter som inneholder termene *fra, frå, ikke, ikkje* og den relative andelen nynorsk (samme metode som i tidligere rapporteringsarbeid)

Nasjonalbiblioteket kommer i samråd med Språkrådet frem til den best egnede beregningsmetoden. Nasjonalbiblioteket mener språkdeteksjon på avsnittsnivå med aggregering på ordnivå (all-institutions-textcat-paragraph-tokens.xlsx) gir det mest realistiske bildet av fordelingen mellom de to målformene. Metoden kan brukes på alle dokumenttyper og tar hensyn til omfanget av dokumentene ved å kvantifisere over ord.

3. Forbehold

Tallene gir etter vår mening et plausibelt bilde av fordelingen av språk på statlige nettsider. Vi vil likevel ta noen forbehold som må tas med inn i det videre arbeidet.

Selv om høstingen har som mål å være så omfattende som mulig, er det visse typer sider som foreløpig ikke er dekket tilstrekkelig. Statsbygg bruker f.eks. JavaScript til å generere alle HTML-sider og har ingen fallback-løsning med statisk HTML. Dette er en utfordring for alle crawlere (inkludert Google). Statsbygg er derfor ikke dekket i denne innhøstingen som man vil se av tabellen. Det kan også være andre institusjoner som bruker utstrakt JavaScript

på enkelte tjenester. Det er et mål i Målfrid-prosjektet å få med så mye dynamisk innhold (altså JavaScript) som mulig, men dette er foreløpig ikke satt ut i drift da det er teknisk utfordrende og krever spesialoppsett for enkeltdomener.

Det er også viktig å nevne at "boilerplate removal" er automatisk og gjort på bakgrunn av bestemte heuristikker og parametere. jusText er kjent for å gjøre en god jobb med dette, men det er samtidig svært sannsynlig at noe relevant innhold er blitt fjernet i prosessen og at noe boilerplate er blitt værende. Det er imidlertid lite sannsynlig at dette får en signifikant påvirkning for resultatene for de enkelte statlige virksomhetene.

Det må også tas forbehold om at språkdeteksjon på korte tekstsegmenter er problematisk, særlig mellom språkformer som i utgangspunktet er svært like. I pilotstudien er dette tatt høyde for i rapportering på avsnittsnivå, ved at avsnitt som er kortere enn 25 ord lukes ut. Grenseverdien er arbitrær, men anvendes likt for alle virksomheter som det rapporteres for.

Likevel vil materiale fra nettet alltid inneholde en betydelig grad av støy. Det er også verdt å nevne at selv om et dokument ligger på et bestemt domene, behøver ikke det å bety at det er offisielt materiale fra en bestemt institusjon. Det kan også være innhold fra tredjepart, f.eks. dokumentasjon. Å skille dette fra egenprodusert innhold, er svært utfordrende.

4. Veien videre

Innhøstingen for 2020 settes i gang i midten av desember 2020 og vil gå frem til midten av januar 2021. Erfaringen fra tre tidligere innhøstingsrunder tilsier at innhøstingen av statlige institusjoner med vår konfigurasjon og innhøstingsdybde tar ca. en måned. Prosesseringen av de innhøstede dataene i etterkant tar igjen ca. to uker. Det er dermed realistisk at Språkrådet vil kunne motta grunnlagsdataene i månedsskiftet januar/februar 2021. Neste innhøsting vil begynne i midten av juni 2021 og gå frem til midten av juli 2021 og resultatene foreligge i månedsskiftet juli/august.

Det vil være nødvendig å gjøre justeringer i domeneliste og unntak for innhøsting underveis. Nasjonalbiblioteket og Digitaliseringsdirektoratet ser på muligheten av å lage et åpent datasett av statlige domenelister, som også kan brukes i Målfrid-arbeidet. Arbeidet med «boilerplate removal» og Språkrådets manuelle filtrering av nettsider vil bli brukt videre i prosessen med å utelukke «irrelevant materiale», slik som database- og katalogdata og innhold av intern karakter.

Leveransen til Språkrådet vil i utgangspunktet bestå av regneark med aggregert data for hver enkelt institusjon. Nasjonalbiblioteket legger samtidig til rette for at Språkrådet kan styre en del av løsningen selv, ved å legge til regler for sider og deler av sider som ikke skal telles (korpusbygging). En tidlig versjon av denne løsningen er allerede presentert for Språkrådet og vil foreligge i en første versjon ved første rapporteringstermin i månedsskiftet januar/februar 2021. Løsningen vil gjøre det mulig å generere egne regneark over et subsett av materialet.