Automatiseret fremfinding af tidligere administrative afgørelser ved hjælp af store sprogmodeller: Erfaringer fra forskningsprojektet LEGALESE (2020-2024) (https://jura.ku.dk/ciir/english/research/legalese-danish-language-processing-for-legal-texts/)

Den moderne velfærdsstat er karakteriseret ved en stor offentlig administration der hver dag træffer tusindvis af afgørelser om borgernes ret til forskellige velfærdsydelser eller forpligtelser til at betale skat. Gennem de seneste 10-15 år har der været et stigende fokus på at kunne effektivisere og forbedre den sagsbehandling der ligger bag disse afgørelser med brug af forskellige former for automatisering.

På nogle områder har automatisering været i brug længe. Det gælder eksempelvis i forhold til fastsættelse af indkomstskat for lønmodtagere og for afgørelser om SU til studerende. På disse områder anvendes regel-baseret computerprogrammer, som er fuldt ud transparente og som kan anvendes til at varetage fuldautomatiseret beslutninger.

På andre områder, som er mere komplekse, har denne slags computerprogrammer vist sig i praksis at volde problemer. Det gælder eksempelvis skattevæsenets gældsinddrivelsessystem EFI (https://www.dr.dk/nyheder/penge/skandalen-om-skats-it-system-efter-11-aars-fiasko-koster-det-millioner-lukke-efi-ned) og senest den automatiserede ejendomsvurdering (https://www.dr.dk/nyheder/penge/har-kostet-milliarder-af-kroner-nu-viser-interne-papirer-et-system-af-forbloeffende) for at beregne ejendomsskat. Manglende hensyntagen til kompleksitet i lovgrundlaget, problemer med retvisende data, manglende inddragelse af berørte borgere mv. kombineret med en ambition om at gennemføre en fuldautomatisering af afgørelser der hidtil har været gennemført manuelt (altså det højest mulige ambitionsniveau for automatisering) er blandt grundende til at disse systemer ikke fungerer optimalt.

Indenfor de seneste 5-7 år er fokus i stigende grad flyttet fra regel-baseret computerprogrammer til de såkaldte dybe neurale netværk og senest store sprogmodeller, som er baseret netop på sådanne netværk. I 2018 offentliggjorde Google deres første store sprogmodel BERT og i 2022 offentliggjorde Microsoft/Open-AI deres interaktive sprogmodel ChatGPT. Lanceringen af disse nye sprogmodeller gav håb om et nyt gennembrud for kunstig intelligens og teknologien har hurtigt vundet frem i mange dele af samfundet. Spørgsmålet er om teknologien også er moden til at understøtte juridisk sagsbehandling. Dette spørgsmål er blevet undersøgt i forskningsprojektet LEGALESE, der i 2020 modtog en bevilling fra innovationsfonden til at undersøge dette.

1. Målsætning og potentiale

Ambitionen med LEGALESE projektet var ambitionen om at bygge en søgemaskine af høj kvalitet til danske administrative afgørelser i den offentlige forvaltning med relevante funktionaliteter, der ville kunne gøre det nemmere for juridiske sagsbehandlere at administrere, navigere og udtrække information fra store mængder af relaterede og semi-relaterede juridiske dokumenter (tidligere afgjorte sager). Ideen var at den nye sprogteknologi (på det tidspunkt primært BERT baseret) ville kunne række ud over traditionelle søgefunktioner baseret på enkeltord, boolske operatorer og trunkering. Ved at bruge BERT-baseret semantisk søgning, ville projektet skabe grundlag for at en sagsbehandler, som skal

afgøre en ny sag, automatisk ville få fremvist de mest lignende tidligere afgørelser truffet af myndigheden.

Fremfinding af tidligere lignende sager rummer en række værdier for både den enkelte sagsbehandler og for myndigheden som helhed. For det første vil et sådant system kunne understøtte princippet om lighed for loven. Ved at fremfinde tidligere afgørelser i sager der ligner, kan det abstrakte princip om lighed for loven konkretiseres i kontekst af en specifik afgørelsespraksis.

En anden vigtig gevinst ved et sådant system består i at mange offentlige myndigheder oplever en betydelig gennemstrømning af medarbejdere. Sagsbehandlere bliver forfremmet og flytter til en anden afdeling, går på barsel og/eller forældreorlov, finder et andet job udenfor myndigheden eller forlader deres arbejde af andre grunde. Ofte skal nye medarbejdere eller vikarer derfor oplæres. Dette kan tage lang tid og kræve mange ressourcer og selv efter endt oplæring vil mange nye medarbejdere opleve en vis usikkerhed fordi retsområdet og myndighedens praksis endnu ikke "sidder helt under huden". I disse situationer kan der være hjælp at hente i at kunne se hvordan tidligere lignende sager er afgjort.

En tredje fordel ved automatiseret fremvisning af tidligere afgørelser er at det kan fremme effektivitet i afgørelsesarbejdet – altså hurtigere sagsbehandling. Ved at få nem adgang til tidligere afgørelser i lignende sager, kan man hurtigere identificere de ræssonementer der er bærende i den pågældende sagstype og ad denne vej vurdere om den nye sag kan afgøres på samme præmisser eller om der er elementer i sagen der skiller sig så meget ud at sagen skal vurderes anderledes.

2. Etiske overvejelser i forbindelse med projektet

Implementering af automatiseret fremfinding og fremvisning af tidligere lignende afgørelser rejser en række etiske spørgsmål, som LEGALESE også har skullet forholde sig til. For det første har det været afgørende for projektet at et sådant system ikke skulle bruges til at automatisere afgørelser. Det er således helt afgørende ved brugen af sådanne systemer at sagsbehandlerne ikke fratages deres mulighed og ansvar for at foretage de fornødne individuelle skøn ved vurderingen af hvordan den enkelte konkrete sag skal afgøres og begrundes.

I forlængelse heraf er det afgørende at et sådant system implementeres på en måde så såkaldt automatiseringsbias (altså tendensen til mere eller mindre blindt at følge en anbefaling fra systemet). Her er den manuelle optræning af juridisk viden indenfor det pågældende domæne gennem orientering i vejledninger, praksis redegørelser, mv. afgørende for at sagsbehandleren kan forholde sig kritisk til systemet. Det er også afgørende at sagsbehandlere ikke afkræves forklaring på om og i hvilket omfang de bruger eller ikke bruger de tidligere sager der fremvises i systemet. Sådanne systemer skal fremstå som et hjælpemiddel der kan lette arbejdet og ikke som et system hvis brug påtvinges sagsbehandleren.

Herudover er det også væsentligt at sådanne systemer ikke skaber eller forstærker eksisterende skævheder i afgørelsespraksis. Netop fordi et af formålene er at fremme lighed for loven er dette vigtigt. I LEGALESE projektet gennemførte vi anonymisering af de tidligere afgørelser der blev indlæst i systemet ved brug navnegenkendelsesteknologi. Ved at

fjerne personnavne, kommunenavne, mv. kunne vi sikre os at navnene ikke kunne fungere som datasammenligningspunkter ved fremfinding af tidligere lignende afgørelser. Dette fjerner ikke helt enhver form for mulig bias i sammenligningsalgoritmen, men minimerer muligheden for sådan bias.

Endelig har det været afgørende at de beslutninger der blev truffet med brug af systemet kunne forklares for slutbrugerne. Da selve fremfindingen af tidligere afgørelser sker automatisk ved brug af en stor sprogmodel (BERT), som er en såkaldt sort boks teknologi, kan selve udfindingen af tidligere afgørelser ikke forklares logisk. Det er dog muligt at forklare i overordnede termer hvordan BERT virker. BERT er et akronym der står for Bidirectional Encoder Representations from Transformers og teknologien er forklaret i en artikel der blev offentliggjort sammen med modellen. Man kan altså ikke logisk forklare hvorfor afgørelse X matches til sag A, men man kan overordnet set forklare hvordan BERT bearbejder sprog og hvilken type beregninger der ligger til grund for at producere de vektorer der sammenlignes og som fører til at afgørelse X beregnes som "mest lig" sag A. Vigtigere er imidlertid at sagsbehandlerne ved gennemlæsningen af de fremfundne afgørelser selv vurderer om de ræssonementer der er knyttet til den tidligere afgørelse er tilstrækkelig lig den nye sag til at de kan overføres og finde anvendelse også i den nye sag. Netop fordi der er tale om et hjælpemiddel til at navigere i institutionens samlede praksis og ikke et system der automatisere afgørelser, denne praksis både etisk og retligt forsvarlig. Fremfinding af tidligere afgørelser med henblik på at uddrage overførbar viden er i øvrigt ikke en ny ide. Det er almindelig kendt at sagsbehandlere ofte orientere sig i myndighedens tidligere afgørelsespraksis med henblik på at finde inspiration til hvordan nye sager skal afgøres. Denne praksis har hidtil ikke været formaliseret på nogen måde, ligesom der heller ikke findes retlig regulering, der foreskriver hvordan sådan fremledning af tidligere praksis skal foregå.

3. Teknisk tilgang

LEGALESE har søgt at indfrie de potentielle fordele ved brug af ny teknologi at bevæge sig ud over state-of-the-art (i 2019, hvor ansøgningen blev skrevet) på tre måder:

(1) Ved at anvende flersproget læringsstrategier, der skulle sætte projektet i stand til at udnytte viden fra ressourcer på flere sprog, og lukke performance hullet mellem dansk og de førende højressourcesprog (særligt engelsk og fransk). Dette foregik ved at gentræne BERT på juridiske EU tekster. Lovgivning og retspraksis fra EU findes på alle EU landes officielle sprog og danner dermed udgangspunkt for træning af en sprogmodel der både genkender juridiske termer og som kan fungere godt på lavressourcesprog (herunder dansk).

(2) Ved at anvende fælles sam-reference opløsning og semantisk opdeling for at opbygge strukturer på dokumentniveau og indlæring af graf-repræsentationsalgoritmer til at generalisere på tværs sådanne strukturer. Dette skulle sikre højere tilbagekaldelse og mere specifik dokumenthentning

(3) Ved at implementere sådanne dokumenthentningsmodeller ved hjælp af indlæringsalgoritmer, der balancerer præcisions-genkaldelse afvejet overfor et mål om at være upartisk med hensyn til demografiske variabler såsom alder, køn, geografisk placering

og indkomst har det været hensigten at undgå uønsket bias i fremfindingen af tidligere lignende sager.

4. Generelle udfordringer

Sammenligning af juridiske dokumenter (afgørelser og sagsakter) ved brug af store sprogmodeller er et meget komplekst område. For det første er dokumenterne i en sag ofte lange hvilket gør dem vanskeligt håndterbare for sprogmodellerne, da flerdimensionaliteten i vektor repræsentationen af dokumenter vokser eksponentielt med længden af dokumentet, hvilket kan føre til manglende retning og præcision i sammenligningsrummet.

For det andet skal der træffes valg om hvilke dokumenter i en ny sag der bedst repræsenterer den juridiske problemstilling der skal tages stilling til. I nogle typer af sager kan dette være enkelt. Foreligger der en simpel ansøgning, måske endda i form af en udfyldt formular kan dette være enkelt. I andre typer af sager består en sag af mange forskellige dokumenter, som alle rummer oplysninger der er eller kan være mere eller mindre relevante for sagens afgørelse. Her skal der træffes afgørelser om alle dokumenter skal medtages og i givet fald med hvilken vægt. Hvis ikke alle dokumenter skal medtages, hvilke skal da medtages og med hvilken vægt. Udfordringen er at disse valg skal træffes generelt for alle sager og i forbindelse med udviklingen af systemet.

For det tredje er det som en sprogmodel betragter som semantisk "lignende" ikke det samme som hvad en jurist anser for at være juridisk lignende. Det er altså afgørende at de tidligere afgørelser der fremsøges er "relevant lignende", hvor relevans afhænger af den juridiske kontekst. Fordi sprogmodeller ikke "forstår" jura, kan det være vanskeligt at træne modellerne til at "se" hvordan den juridiske kontekst påvirker værdien af den semantiske lighed mellem to dokumenter som modellen udfinder.

5. Udfordringer specifikt for LEGALESE

Udgangspunktet for LEGALESE projektet var et ønske om at bidrage med fremfinding af afgørelser fra lignende sager på et område med udgangspunkt i et område med et oplevet behov. Efter dialog med Ankestyrelsen udvalgte vi afgørelser af sager, hvor borger har fået afslag fra deres kommune på hjælp efter servicelovens §41 eller §42 (nu barnets lov §86 og §87)[1] og hvor de har klaget til Ankestyrelsen over kommunens afgørelse. Vi etablerede derfor en database bestående alle Ankestyrelsens tidligere afgørelser efter servicelovens §41 og §42 (§86 og 87 i barnets lov) og etablerede en automatisk opdateringsfunktion, således, at de nyeste afgørelser også kom til at indgå i databasen. Ved brug af software som blev stillet til rådighed af KMD og tilpasset til projektet, kunne sagsbehandleren med udgangspunkt i en åbnet sag, se hvilke tidligere afgjorte sager der bedst matchede den åbnet sag. "Bedste match" var de sager som den anvendte sprogmodel[2] definerede som mest lignende.

---

[1] Se lovteksterne her: https://www.retsinformation.dk/eli/lta/2024/890

[2] Den anvendte sprogmodel var specifikt udviklet til projektet af forskere fra Københavns Universitet, Datalogisk Instituts Coastal gruppe gennem en tilpasning af BERT. Modellen er tilgængelig her: https://github.com/coastalcph/danish_legal_lms

Offentlig sagsbehandling på det område vi udvalgte i projektet, tager udgangspunkt i den enkelte borgers samlede situation. Kernen i §41/§86 sagerne er spørgsmålet om hvad der i given sag kan siges at udgøre ”nødvendige merudgifter ved forsørgelse i hjemmet” for en familie med et barn der har ” betydelig og varigt nedsat fysisk eller psykisk funktionsevne eller indgribende kronisk eller langvarig lidelse”. En betingelse for at få hjælp er at ”merudgifterne er en konsekvens af den nedsatte funktionsevne eller den indgribende kroniske eller langvarige lidelse.”

Dette betyder at sagsbehandlerne skal skønne over om en konkret udgifter er ”nødvendig” og om udgiften er en ”konsekvens af den nedsatte funktionsevne”. Der afgøres ca. 800 sager af denne slags i Ankestyrelsen hvert år og de udgifter der søges dækket gennem denne bestemmelse varierer ganske betydeligt (fra udgifter til diverse former for terapi og kognitiv støtte over udgifter til særlige madvarer, møbler, tekniske hjælpemidler, til transportudgifter, anlæg af særlige adgangsforhold, mmm). Ved afgørelsen af om der skal tildeles hjælp er det ofte en kombination af karakteren af den nedsatte funktionsevne, sammenhængen mellem dette og det hjælpemiddel eller den service som udgiften dækker over og en vurdering af om udgiften er en sædvanlig udgift eller om udgiften er affødt af den nedsatte funktionsevne. Ofte skal der foretages et komplekst samlet skøn for at kunne foretage en juridisk forsvarlig afgørelse.

De samme faktiske forhold kan derfor have forskellig betydning i forskellige situationer, fordi konteksten er forskellig. Dette kan sprogmodeller pt. ikke tage højde for, når de bruges til at fremlede tidligere sager fordi modellerne ikke har forudsætninger for at vurdere hvad der er ”sædvanligt” eller hvad der er ”nødvendigt” i en given kontekst. Dette har i projektet manifesteret sig ved at det særligt er i meget komplekse sager, at sprogmodellerne har vanskeligt ved at finde tidligere sager der af sagsbehandlerne opleves som sammenlignelige. Den juridiske vurdering af hvad der udgør et relevant administrativt præjudikat opfanges med andre ord ikke særlig nemt af sprogmodellerne.

Tilsvarende udfordringer oplevede projektet i sagerne om lønkompensation (§42/87). I disse sager skal Ankestyrelsen vurdere om det er en nødvendig konsekvens af den nedsatte funktionsevne, at barnet eller den unge passes i hjemmet, og at det er mest hensigtsmæssigt, at det er moderen eller faderen, der passer barnet eller den unge (fremfor at barnet eller den unge passes i en institution eller af en professionel hjælper). Igen er det den meget konkrete vurdering af forældrenes arbejdsmæssige situation kombineret med den konkrete funktionsnedsættelse og familiens situation samlet set der er afgørende for om der bevilges kompensation for tabt arbejdsfortjeneste.

6. Tekniske tilgange og test

Henover projektets levetid afprøvede vi forskellige løsninger i forhold til at fremfinde sammenlignelige sager. I første runde udviklede vi en algoritme der læste 4 forskellige dokumenter i hver sag: 1) Kommunens afgørelse, 2 borgerens klage, 3) kommunens revurdering, 4) Ankestyrelsens afgørelse. Hen over disse 4 dokumenter beregnede vi en vektorprofil. For hver allerede afgjort sag (i alt ca. 10.000 afgørelser) beregnede vi en sådan profil. For hver ny sag beregnede vi en vektorprofil baseret på 3 dokumenter: 1) Kommunens afgørelse, 2 borgerens klage, 3) kommunens revurdering. Herefter sammenlignede vi den nye sag med alle de afgjorte sager og fremviste herefter de mest lignende afgjorte sager til

sagsbehandleren. Til at læse og profilere de gamle sager anvendte vi ovennævnte sprogmodel (se note 2). Herefter testede vi resultaterne i sagsbehandlingspraksis ved at lade 4 sagsbehandlere teste algoritmen på i alt 60 sager. Testforløbet blev gennemført ved brug af et IT system udviklet af KMD kaldet Case Insight[3].

7. Resultater og justeringer i tilgang

7.1.Første testrunde

Resultatet af denne første runde test var at sagsbehandlerne oplevede de fremfundne sager som ikke sammenlignelige. Resultaterne fremstod for sagsbehandlerne som mere eller mindre arbitrære. I enkelte tilfælde var der brugbare match, men sagsbehandlerne oplevede dette som tilfældigheder og kunne ikke se nogen systematik i de fremfundne "match". Testerne havde adgang til og undersøgte ikke kun de sager som havde den højeste "ligheds-score", men også de næstehøjeste, tredjehøjeste osv. Sagsbehandlerne kunne også se om de fremfundne sager var 41/86 eller 42/87 sager og kunne selv vælge om de ville springe de sager over der ikke skulle behandles efter samme bestemmelse.

7.2.Anden testrunde

På baggrund af de dårlige resultater fra denne første runde, besluttede vi i anden runde at gennemføre en test med TF/IDF som er en grundlæggende og velkendt sprogteknologi, der er fuldt ud transparent. Den grundlæggende tankegang bag TF/IDF er er at dokumenter der deler de samme sjældne ord vil have mere til fælles end dokumenter der ikke deler sjældne ord. Sjældenhed beror i denne kontekst på hvor ofte ordet optræder i den samling af dokumenter, hvori sammenligningen finder sted.

Belært af erfaringerne fra første testrunde, hvor interviews med de testende sagsbehandlere viste at sagsbehandlerne oplevede at nogle sager der er blev fremvist som sammenlignelige faktisk var sammenlignelige forstået på den måde at de havde nogle sproglige fællestræk med den nye sag der skulle afgøres, besluttede vi at gennemføre en manuel revision af de ord der skulle indgå i TF/IDF modellen. Denne revision blev foretaget af projektets PI og en medarbejder fra Ankestyrelsen som var tilknyttet projektet og som havde praktisk erfaring med sagsbehandling (men som ikke selv deltog i at teste projektets sammenligningssystem).

Som endnu et led i anden testrunde besluttede vi at eliminere nogle af dokumenterne fra den oprindelige tilgang til sags sammenligning. I lyset af de dårlige resultater fra første runde blev vi efter konsultation med de testende sagsbehandlerne i Ankestyrelsen enige om at fjerne borgernes klageskrivelser og kommunens revurdering, således at vi kun beholdt den oprindelige kommunale afgørelser og Ankestyrelsens afgørelse som sammenligningsgrundlag.

---

[3] En introduktion til systemet findes her: https://publikationer.kmd.dk/data-og-analyse/kmd-case-insight/kmd-case-insight/?page=1

Begrundelsen for dette var en antagelse om at borgernes klageskrivelser, fordi de havde en meget ujævn karakter (nogle klager var skrevet af borgerne selv og uden kendskab til loven – andre var skrevet af advokater på vegne af borgerne; nogle klager var meget korte (få linjer) – andre klager var meget lange (10-20 siders tekst). Samlet set var det således vores vurdering at disse klageskrivelser introducerede en del "støj" (irrelevant tekst) i forhold til opgaven med automatiseret fremfinding af sammenlignelige sager. Som en konsekvens af at fjerne klageskrivelserne, fjernede vi også kommunens revudering, da disse dokumenter i vid udstrækning responderer på klageskrivelserne. De afgjorte sager blev herefter repræsenteret alene ved kommunens oprindelige afgørelse i sagen samt ankestyrelsens afgørelse.

Endelig gennemførte vi en automatiseret anonymisering af dokumenterne i de gamle sager. Sagsbehandlerne havde i første test runde observeret at fremfundne sager ofte kom fra samme kommune, hvilket førte til en mistanke om at der indsneg sig uønsket bias i den automatiseret fremfinding (kommunenavn er ikke en relevant oplysning for sagsbehandlingen). Vi anvendte derfor et NER (Named Entity Recognition) værktøj til at fjerne alle stednavne og lignende i dokumenterne i de gamle sager.

Efter at have gennemført disse tilpasninger gennemførte vi endnu en test-runde med 60 sager. Resultatet af disse test viste en klar forbedring i forhold til første runde, men systemet kunne stadig ikke konsistent fremvise relevant sammenlignelige sager. Sagsbehandlerne udtrykte i denne runde at de oftere fik fremvist relevante sager, men der var ikke konsistent gode resultater og ofte var det i sager, som i forvejen var relativt enkle at behandle at systemet fremviste relevante tidligere afgørelser. De enkelte sagsbehandlere oplevede resultaterne lidt forskelligt – positive indtryk af systemets evne til at fremvise relevante tidligere afgørelser varierede fra 0.0 – 0.5, men som nævnt var disse i overvejende grad knyttet til nye sager, som oplevedes som "nemme sager" hvorfor der ikke var noget stort behov for at læne sig op af / hente inspiration fra tidligere praksis.

7.3.Tredje testrunde

I tredje og sidste testrunde vendte vi tilbage til udgangspunktet for projektets ambition, nemlig at implementere en stor sprogmodel i et system til fremfinding af administrativ praksis. Vores analyse var at den BERT baseret sprogmodel der var udviklet i projektets første fase var trænet på juridisk tekst (EU lovgivning og EU domme), som nok havde givet modellen et større juridisk ordforråd, men som ikke havde indfanget de sproglige nuancer som var relevante for de afgørelser vi testede på[4]. Vores tilgang i denne sidste runde var, baseret på feedback fra sagsbehandlernes erfaringer, at der var brug for en sprogmodel, som var mere forankret i almindeligt dansk sprogbrug. Vi valgte derfor at bruge en model[5], som, i sammenligning med andre sprogmodeller, udviste en høj præstationsgrad[6]. Samtidig justerede vi endnu en gang på parametrene for de gamle sager, således at vi i denne sidste runde alene sammenlignede op imod Ankestyrelsens egne afgørelser og udelod de

[4] Det var ikke muligt at træne sprogmodellen på den store mængde af tidligere afgørelser fordi disse rummer sensitive personlige data. Den oprindelige tilgang var derfor blot at træne en sprogmodel på et korpus af juridiske tekster, for at fremme modellens evne til at sammenligne juridiske afgørelser. Dette viste sig ikke at være en brugbar løsning.

[5] https://huggingface.co/intfloat/multilingual-e5-large

[6] https://kennethenevoldsen.github.io/scandinavian-embedding-benchmark/

underliggende dokumenter i sagerne (kommunens oprindelige afgørelse, klageskrivelsen, osv).

Resultaterne fra denne sidste testrunde lå nogenlunde på højde med det vi oplevede i anden testrunde, men denne gang altså med brug af en stor sprogmodel. Igen var der stor spredning i den oplevede brugbarhed af de fremfundne sager. Enkelte sagsbehandlere oplevede at systemet ikke gav nogen brugbare fremvisninger af tidligere sager, medens andre oplevede at fremfundne sager i flere tilfælde indeholdt "brugbare momenter", det vil sige begrundelsestekst der knytter sig enkelte elementer i afgørelsen, f.eks. i forhold til §41/86 sammenligningsgrundlag (hvad er almindeligt forekommende udgifter og dermed ikke en udgift der er en "konsekvens af den nedsatte funktionsevne") eller nødvendighedsvurdering[7]. Også i sager om tabt arbejdsfortjeneste (§42/87) er der flere eksempler på at konkrete sagsforhold, f.eks. deltagelse i møder om børns trivsel i skole og/eller institution og sager om fastlæggelse af løngrundlaget for tabt arbejdsfortjeneste samt sager om ændringer i allerede bevilget tabt arbejdsfortjeneste. Samlet set ligger den oplevede brugbarhed af de fremsøgte sager igen mellem 0.0 og 0.5 og igen er det den generelle opfattelse blandt sagsbehandlerne at det primært er i de "nemme" sager, hvor der findes de bedste "match".

En enkelt sagsbehandler oplever at de fremfundne sager i mange tilfælde er "gamle", hvilket af den pågældende sagsbehandler defineres som sager fra perioden 2018-2021. Dette tyder på at udviklingen i praksis på det udvalgte sagsområde (§41/86 og §42/87) går relativt hurtigt uanset at lovteksten er den samme som ved den oprindelige servicelovs vedtagelse i 2005. Dette kunne tyde på at der er potentiale for et selvstændigt forskningsprojekt om udviklingen i Ankestyrelsens afgørelsespraksis på området.

7.4.Konklusion på test

Samlet set lykkedes det ikke at udvikle eller finjustere en sprogmodel som konsistent fremfinder juridisk set relevant tidligere afgørelser. Ambitionen om at understøtte højere effektivitet og konsistens i sagsbehandlingen kunne således ikke indfries. Selv om der var positive oplevelser med systemet var disse for spredte til at systemet ville kunne implementeres som et egnet hjælpemiddel i Ankestyrelsens sagsbehandling på §41/86 og §42/87 området. Dette giver naturligvis anledning til refleksion over hvordan man vil kunne udvikle et sådant system, så det i højere grad fremfinder tidligere afgørelser der opleves som juridisk relevante og brugbare som understøttelse af sagsbehandlingen i nye sager.

8. Vigtigste opgaver og udfordringer i forhold til videreudvikling af LEGALESE

Administrativ sagsbehandling og administrative afgørelser i den offentlige forvaltning kan være meget omfattende og hver enkelt sag kan rumme mange forskellige problemstillinger, hver med flere forskellige retlige dimensioner knyttet til forskellige faktuelle oplysninger i sagen. Baseret på det omfattende testarbejde, som er blevet udført i kontekst af LEGALESE kan det konkluderes at Afgørelsesdokumenter – i hvert fald på det sagsområde, som LEGALESE har testet på – udgør en *for* kompleks juridisk enhed for både enkeltords-baseret sprogmodeller og transformerbaseret (store) sprogmodeller. Sprogmodellers beregning af hvad der udgør mest lignende dokument fanger med andre ord ikke nødvendigvis den

[7] Se om lovgrundlaget og dermed bedømmelsestemaerne ovenfor i afsnit 5 med henvisninger til lovteksten.

juridiske forståelse af hvad der er mest lignende. Der kan være forskellige forklaringer på dette. En oplagt forklaring er at der simpelthen ikke findes tidligere afgørelser der ligner i juridisk forstand. Selvom der er findes 10.000 tidligere afgørelser om §41/86 og §42/87 som man kan sammenligne med kan det godt være at der ikke er nogen af disse som opfattes som "lignende" i juridisk forstand set i forhold til den nye sag som er under behandling. Det forekommer dog usandsynligt at den lave grad af oplevet lighed, der blev observeret i vores test skulle være et resultat af at der ikke findes juridisk "lignende" tidligere afgørelser blandt en så stor mængde af afgørelser der vedrører en så afgrænset problemstilling. Det er i hvert fald sandsynligt at det vil være muligt at opnå bedre resultater ved at videreudvikle og raffinere tilgangen til automatiseret fremfinding af tidligere afgørelser. Her skal der sættes fokus på fire tiltag, som vurderes at kunne skabe bedre resultater:

1) Anvendelse på et mindre komplekst retsområde
2) Strukturering og opdeling af afgørelser i afsnit
3) Anvendelse af forstærkningslæring gennem menneskelig tilbagemelding

Ad 1)

LEGALESE projektet blev udviklet med en meget ambitiøs målsætning for øje – nemlig at understøtte administrative juridiske afgørelser af enhver art i den offentlige administration ved at fremsøge tidligere lignende afgørelser. Projektafviklingen har vist at juridisk sagsbehandling på nogle retsområder (in casu servicelovens/barnets lov §41/86 og §42/87), hvorfor der kan være fordel i at udvikle og afprøve denne form for understøttelse på et mindre komplekst retsområde for at indhente yderligere erfaringer.

Ad 2)

Afgørelsesdokumenter kan rumme en længde og en kompleksitet som indebærer at det som er sprogligt "lignende" kan være vanskeligt at beregne på en klar og entydig måde. Fordi det kræver en juridisk vurdering at identificere den tekst der er "relevant lignende" kan de fremfundne resultater fremstå som arbitrære. En måde at gøre det nemmere for sprogmodellerne at identificere hvad der "relevant lignende" i juridisk forstand vil være at gennemføre sammenligner af mindre tekstbidder. Flere sagsbehandlere i vores test oplever som også nævnt ovenfor, at de fremfundne sager/afgørelser rummer brugbare "momenter" eller "elementer". Dette indikerer at det ikke så meget er hele sager eller hele afgørelser, men snarere dele af sager og afgørelser som er rette niveau for brugbar understøtte af sagsbehandlingen. Fremfor at sammenligne op imod hele afgørelsesdokumenter kunne man opdele afgørelserne i afsnit. Tilsvarende kunne sagsbehandlerne udvælge relevante afsnit eller passager fra den nye sag og markere disse med henblik på at matche disse til et afsnit i en tidligere afgørelse. Dette vil alt andet lige fremme sandsynligheden for et relevant match.

Ad 3)

Det er ikke altid at en søgefunktion virker optimalt umiddelbart efter at den er udviklet. Som en måde at fremme søgefunktioners virkemåde, kan man installere et feedback loop, som gør det muligt for brugerne af søgefunktionen at sende information tilbage til den underliggende

sprogmodel, som så justerer på sine parametre i forhold til den modtagne feedback. Da det i vidt omfang sagsbehandlernes vurdering af hvad som er juridisk relevant som udgør målet for hvad en teknologi skal kunne understøtte vil det være relevant løbende at optimere en søgefunktion ved bruge feedback til at gøre modellen bedre til at identificere brugernes forståelse af hvad der er gode/relevante henholdsvis dårlige/irrelevante søgeresultater.

9. Andre udviklinger i projektperioden

Selv om den forskningsmæssige del af projektet i høj grad er indfriet har der i perioden været betydelige forskningsgennembrud andre steder fra, som har sat projektets sprogteknologiske ambition lidt i skyggen. Udrulningen af ChatGPT i slutningen af 2022 og de efterfølgende konkurrerende store sprogmodeller (eksempelvis Claude, Llama og Mistral) og den videre udvikling af GPT fra 3,5 til GPT4 og senest GPT4o har fået meget stor opmærksomhed og har flyttet state-of-the-art på en måde som det ikke var muligt at forestille sig i 2019, hvor BERT udgjorde den videnskabelige for-front i sprogteknologi.

Også lanceringen af EU's kunstig intelligens forordning i projektperioden har haft stor betydning. Forslaget til forordningen blev fremlagt af EU kommissionen i Apil 2021 og forordningen blev vedtaget i begyndelsen af 2024 og er nu trådt i kraft. Forordningen får stor betydning for de krav der stilles til kunstig intelligens på de såkaldte høj risiko områder, som bl.a. omfatter AI-systemer, der tilsigtes anvendt af offentlige myndigheder i forbindelse med afgørelser om ret til offentlige bistandsydelser og -tjenester. Altså en meget stor del af den offentlige velfærdsstats administration.

Disse to meget store ændringer på projektets forskningsområde har haft betydning for forskningen på den måde at interessen har rykket sig mere i retning af potentialet for at generativ kunstig intelligens og et stigende fokus på at sikre compliance med AI Act.