

Skema: Ansøgning om resultatkontraktmidler 2019-2020

Institut(ter): Alexandra Instituttet	Aktivitetsplan (titel): Dansk for alle Idéforslagstitel på bedreinnovation.dk: Dansk for alle	Aktivitetsplan nr.: 2	FoU
1) Manchettekst (kort resumé)			
<p>Der er nye muligheder for automatisk håndtering af tekst til næste generation af services inden for sproget dansk. Der er dog mangel på viden og adgang til ressourcer både i form af anoterede datasæt men også i form af metoder og værktøjsbyggekodser.</p>			
2) Aktiviteten kort (resumé)			
<p>Vi ønsker at udvikle et sæt danske ressourcer til automatisk sproghåndtering. Dvs. et sæt sprogressourcer, der tillader SMV'er (og alle andre) at udnytte automatisk håndtering af det skrevne danske sprog i deres produkter og deres daglige arbejdsgange. Ressourcerne tager form af datasæt, metoder, regelsæt, værktøjer, opskrifter og tilgængelige best practices, så virksomheder nemt kan komme i gang med at bygge kunstige intelligenssystemer, der håndterer dansk tekst. Der vil være tale om et sæt af byggekodser, der kan anvendes på tværs af forskellige virksomheder og forretningsdomæner.</p> <p>I projektets løbetid vil fokus være på indsamling af komponenter, udstilling i et repositorium og basisydelser baseret på disse komponenter. Der er altså tale om en "RedHat" kommerciel model, hvor vi stiller basiskomponenter til rådighed, og hvor teknisk support og specialkomponenter kan tilkøbes på kommercielle vilkår. Vi vil demonstrere de udviklede ressourcer i cases med særligt udvalgte virksomheder og vise, hvordan man kan digitalisere og automatisere håndtering af skrevet dansk tekst ved hjælp af kunstig intelligens.</p> <p>Baseret på disse ressourcer og kompetencer opbygget i denne aktivitetsplan vil Alexandra Instituttet tilbyde en række af ydelser til virksomheder, der ikke selv ønsker af opbygge kompetencer samt til virksomheder med særlig komplekse udfordringer. Der er tale om ydelser, der udbydes bredt til alle virksomheder, som arbejder med automatisering og digitalisering af dansksprogede tekster. Der er således ikke en kundegruppe segmenteret efter fag men derimod segmenteret efter behov (behovet for at kunne automatisere behandlingen af dansksprogede tekster). En lang række virksomheder har meldt sig som interessenter og som potentielle casevirksomheder. Det spænder fra virksomheder, der ønsker at automatisere dele af regnskabs- og revisionsbranchen, som f.eks. Digital Revisor, til branchenetværk som f.eks. Iværksætter Netværk (IVN) til store virksomheder som IBM, ATP og offentlige institutioner som Erhvervsstyrelsen.</p> <p>Overordnede aktiviteter</p> <ul style="list-style-type: none"> • Viden om bedste basisressourcer til moderne sproghåndtering, specifikt for dansk • Opbygning af sprogspecifikke basisressourcer (datasæt, metoder, løsningsbeskrivelser, opskrifter og best practice beskrivelser. Regelsæt, værktøjskasse, anvendelse af Big Data og særligt deep learning teknologier til sproghåndtering). • Kompetencer omkring håndtering af gængse udfordringer på sprogområdet. Som for eksempel, sentimentanalyse, intentionsanalyse af dialoger, generering af svar i Q&A systemer samt ikke mindst beskyttelse af personfølsomme oplysninger i tekstbaserede systemer. 			

Dansk for alle vil udstille et online tilgængeligt repositorium af komponenter, som alle virksomheder kan benytte til automatisk håndtering af sprog. Målet er, at mange af de tilgængelige komponenter skal være open source/åben adgang, som alle mere eller mindre bare kan tilgå og benytte. Alexandra Institutet vil så udvikle en række af kommercielle services og produkter oven på de åbne komponenter.

3) Markedsbehov, erhvervs- og samfundsmæssige potentialer

Behovet er, at danske virksomheder lige så nemt som udenlandske virksomheder kan tage moderne teknologier i brug, når det drejer sig om at høste ustruktureret, sproglig information.

Alle virksomheder, som arbejder med automatiseret sprogforståelse vil kunne drage fordel af vores åbne ressourcekatalog og ydelser. Som primær målgruppe for vores kommercielle ydelser ser vi SMV'er, som ikke ellers har oplagte ressourcer til at løse denne type af teknologiske udfordringer. Alle virksomheder, der i dag vedligeholder sprogteknologier vil få glæde af nye ressourcer og værktøjer. Det kan være advokatfirmaet, der vil automatisere arbejdsgange eller søge i interne dokumenter, og som i dag har nogle udtræksregler for at finde [navne, adresser osv.](#) Det kunne også være firmaet, som får tekstrapporter, som skal behandles og forstås. F.eks. taksatorrapporter til forsikringselskaber, hvor den strukturerede del af rapporten behandles automatisk, mens [fejl i struktur eller ren tekst i dag kræver manuel håndtering](#).

Også virksomheder, der i dag har fravalgt at bruge automatisk sprogteknologi pga. udfordringerne ved den klassiske tilgang, vil kunne drage fordel af et stærkt sæt af ressourcer til automatisk sproghåndtering. Det kunne f.eks. være [revisoren, der ønsker at automatisere importen af kundekonti og posteringer](#), eller det kunne være virksomheder, der ønsker at udnytte interne dokumentdatabaser og lave vidensdeling på tværs af virksomheden, men som har opdaget, at manuelt vedligeholdte [keywords og metadata ikke leverer varen, og som ønsker en semantisk søgning](#).

Forudgående dialog med målgruppen og relevante innovationsnetværk

Alexandra Institutet har sammen med innovationsnetværket InfinIT arrangeret workshops i anvendelse af dansk NLP i f.eks. inden for kundeservice bl.a. i Aarhus i august i år. Det arrangement og et tilsvarende i København i maj måned er nogle af de meste besøgte workshops i InfinIT med henholdsvis 100 og 180 deltagere. Både antallet af deltagere samt den dialog, vi har haft med virksomhederne før og efter samt kommentarerne på [bedreinnovation.dk](#) vidner om det store og akutte behov blandt danske virksomheder for bedre værktøjer og ressourcer til automatisk håndtering af dansksprogede tekster.

Gartner forudsiger at 25% af al kundesupport i 2020 vil blive foretaget af chatbots, dvs. systemer der automatisk håndterer det skrevne sprog. Dette medfører en effektivisering og digitalisering i kundesupport, som danske virksomheder ikke kan være med i, hvis vi ikke har høj kvalitets værktøjer til at håndtere alle disse kundehenvendelser. Danske virksomheder stilles svagt i forhold til deres internationale konkurrenter, hvis ikke de får adgang til de ressourcer, vi vil udvikle og udstille i dette projekt.

Sammenhæng med nationale strategier og internationale målsætninger

Opgaven med at lette informationsudtagning af tekster på dansk ligger oplagt som en vigtig del af [Strategi for Danmarks digitale vækst](#). Bl.a. i indsatsområderne "SMV: Digital" og "Agil erhvervsrettet regulering", hvor bl.a. sprogteknologi og automatisk håndtering af dansksprogede tekster oplagt giver særligt SMV'erne gode muligheder for at øge deres omsætning og konkurrenceevne.

Vi mener, at dette er en opgave i GTS-regi, da det har samfundsmæssig stor betydning, og fordi der er brug for en investering på tværs af domæner for at kunne udvikle de nødvendige ressourcer. En stor investering, som ikke er økonomisk rentabel for en enkelt virksomhed at finansiere for at udvikle en domænespecifik

sprogløsning. Vi har alle større fordel af ét sæt af højkvalitets ressourcer frem for mange proprietære ressourcer baseret på lille databredde.

Ved at Alexandra Institutet løser ressourceproblemet for dansk tekst og udvikler en serie højkvalitets teknologiske byggeklodser, kan vi sikre, at det bliver nemt for virksomheder at komme i gang med automatisk sproghåndtering, og at fokus for virksomhederne bliver på at udvikle innovative nye services.

4) Videnspredning og inddragelse

Inddragelse af danske virksomheder

Aktivitetsplanen vil starte ud med en afdækning af virksomhedernes behov. Dette kan enten ske gennem observationer, studier, workshops eller lignende med fokus på systematisk behovsafdækning og -analyse af nuværende teknologibehov og anvendelsesmuligheder. Desuden vil virksomhederne løbende blive involveret gennem casearbejde, hvor vi i projektet går sammen med virksomheder om at løse specifikke dele af deres NLP-problemer ved hjælp af de algoritmer, data og opskrifter, der udvikles i projektet. Vi forventer mindst 3 cases i projektets løbetid. Dette bidrager både til at scope projektet og sikre, at der er arbejdes på virksomhedsrelevante problemstillinger i projektet.

Virksomheder inddrages desuden gennem videnspredningsarrangementer så som gå-hjem-møder og workshops. Der etableres også en følgegruppe med deltagelse af 3-5 virksomheder fra målgruppen, som deltager i årlige møder, hvor formålet er at drøfte resultater og status på udviklingen af teknologiske serviceydelser, og hvorledes disse matcher målgruppens behov - herunder eventuelle ændringer i konkurrencesituationen. Der opnås herunder en afklaring af Alexandra Instituttets rolle ift. private rådgivere og leverandører mhp. at optimere synergi og komplementaritet og undgå konkurrenceforvridning.

Samarbejde med eksisterende initiativer

Vi vil samarbejde med [Dansk Sprognævns \(DSN\) arbejdsgruppe omkring moderne sprogteknologi](#). Gruppen er planlagt til at afslutte deres arbejde i løbet af 2018, men deres viden er oplagt vigtig at få indlejret i denne aktivitetsplan, og vi er allerede i dialog. Derudover har vi lavet en videnspredningsaftale med Innovationsnetværket for IT, InfinIT, som i perioden 2019-2020 bl.a. har fokusgrupper omkring kunstig intelligens og big data. Dette samarbejde kan både skabe en platform for at dele viden opnået i aktivitetsplanen men også give mulighed for at komme i bredere dialog med interesserede virksomheder og forskere uden for projektet.

Formidling af aktiviteter og resultater

Dette sker ved afholdelse af videnspredningsarrangementer, hvor vi inviterer bredt og fortæller om opnåede resultater og muligheder – bl.a. som nævnt i samarbejde med InfinIT.

Der vil også ske formidling gennem nyhedsbreve og via Alexandra Instituttets hjemmeside og interessenter vil kunne finde relevant information på vores online ressourcekatalog, hvor vi sætter fokus på interessante resultater, datasæt og tutorials.

Vi forventer at dele den nye viden med i størrelsesorden 60 unikke virksomheder gennem vidensdelingsarrangementer som kurser, seminarer, workshops, conferenceoplæg mv. Derudover forventer vi, at mange virksomheder finder vores nye viden via nyhedsbreve, online casebeskrivelser samt via vores ressourcekatalog.

Publikationer

Vi forventer i alt mindst 4 indlæg i projektets løbetid – f.eks. artikler i fagmedier, videnskabelige artikler, publikationer og conferenceartikler. Samt mindst 2 præsentationer i relevante fora/konferencer.

5) Konkrete aktiviteter

Målet i denne aktivitetsplan er at gøre det nemmere for danske virksomheder at indbygge automatisk dansk sprogbehandling af høj kvalitet i deres produkter. Konkret vil vi arbejde med:

Bygge et repositorium af

- datasæt
- NLP-algoritmer
- opskrifter/tutorials på, hvordan man løser typiske NLP-opgaver
- whitepapers og best practice beskrivelser over NLP-løsningsmodeller, arkitektur og andre valg, man skal være opmærksom på.

Dette repositorium er inspireret af tilsvarende data repositorer, som kendes fra Smart Cities og open data/åben adgang aktiviteter.

Udviklingscases med virksomheder fra målgruppen

Der er i aktivitetsplanen planlagt minimum 3 caseforløb, hvor bl.a. Digital Revisor og ATP har tilkendegivet interesse i at deltage. De ville danne udgangspunkt for fokus og prioritering i projekt. Sammen med casevirksomhederne vil vi udvikle demoer, der løser casevirksomhedernes behov omkring sprogbehandling. Disse demoer vil fungere som demonstratorer og eksempler på fordelene ved at bruge ressourcerne udviklet i denne aktivitetsplan.

Videnhjemtagning og kompetenceopbygning

Målet er at opbygge kompetencer i teknologi og viden samt hjemtage viden om, hvordan konkrete udfordringer i dag løses på f.eks. engelsk. Denne kompetenceopbygning vil blive udstillet som ressourcer i vores katalog i en gratis basisform og i en mere avanceret tilkøbsform.

Opbygning af viden høstet fra samarbejde med forskere ved KU, samt ved at følge det arbejde, der foregår i forskningsmiljøerne omkring at øge kvaliteten af automatisk sprogforståelse.

Vi vil opbygge kompetence både via det forskningsmæssige samarbejde men i endnu højere grad gennem konkret erfaring om problemstillinger og løsninger via af vores caseprojekter.

Videnspredning

Der vil løbende i projektet blive indbudt til og afholdt videnspredningsarrangementer så som workshops, foredrag, sparringsessioner mv. Vi planlægger at publicere og informere både via dagspressen og fagrelevante medier. Derudover planlægges også information via Alexandra Instituttets online publiceringskanaler. Se også milepæle i afsnit 9.

Forretningsudvikling

Målet er, at det skal være nemt for alle at komme i gang, og de, der vil selv, kan få lov at gøre det ved at udnytte den åbne del af ressourcekataloget. De virksomheder, der har brug for mere avancerede løsninger, og som gerne vil tilkøbe support, eller som ikke ser det som deres kerneforretning at være skarpe på NLP og automatisk håndtering af dansk, bliver kommercielle kunder hos Alexandra Instituttet og får mulighed for at tilkøbe de kommercielle ydelser udviklet i denne aktivitetsplan. Vi udvikler og udbyder 3 nye markedsrelevante ydelser til konkrete virksomheder i projektets periode.

Barrierer, uvisheder og risici

Dansk lovgivning for brug af datasæt og ressource til NLP-modeller halter bagud i forhold til f.eks. andre nordiske lande, og som resultat heraf er sproressourcer til dansk af mindre omfang og kvalitet. En barriere for at få nem adgang til store datasæt er lovgivningen, som i højere grad understøtter brug af tekster til læsning (betaling for ophavsret) og i mindre grad brug af tekster til træning af algoritmer, som kun tilgår copyright indhold på en meget indirekte måde.

Bemærk, at der findes mange NLP-ressourcer, der allerede er åbne, og som virksomheder vil kunne drage fordel af i det øjeblik, der bliver lavet nem adgang til ressourcerne for eksempel gennem vores ressource repository, som vil muliggøre et "single-point-of-discovery" af ressourcer for danske virksomheder.

Det skal desuden bemærkes, at det er en lovmæssig barriere/udfordring i stil med dem der kendes fra åbne dataplatforme, som dem Alexandra Instituttet har arbejdet med i mange år. Således er succes af kataloget ikke afhængigt af en lovændring, men en lovændring vil blot øge mængden af tilgængelige ressourcer markant.

Organisatoriske, ledelsesmæssige og bemandingsmæssige overvejelser

Denne kritiske opgave med at indsamle og gøre relevant information og ressourcer omkring automatisk sproghåndtering lettilgængelig, er en udfordring, der kendes andre steder fra. For eksempel opstod en lignende udfordring i 80'erne og 90'erne, hvor regelbaseret talegenkendelse fik et gennembrud, og der var brug for viden og ressource til talegenkendelse. Dengang resulterede det i, at mange forskellige virksomheder genererede deres egne ressourcer og datasæt med det resultat, at ingen af produkterne levede op til det fulde potentiale.

Der har i forskningsverdenen af flere omgange været projekter, hvor information og ressourcer til automatisk sproghåndtering er blevet genereret, indsamlet og præsenteret. Desværre stoppede disse projekter alle efter videnskabelig publikation af resultaterne.

På baggrund af disse udfordringer står GTS-nettet i en særlig stærk position til at adressere den ressourcemæssige udfordring omkring automatisk sprogforståelse. Da vores mål er at hjælpe danske virksomheder bredt, vil vores repository ikke stoppe med at vokse, når en enkelt delopgave er løst men derimod, når vi har sikret stærke generaliseringsmuligheder for de modelbaserede ressourcer. Samtidigt er vores fokus ikke forskningsdrevet, og vi kan opretholde et repository så længe, det har værdi for de danske virksomheder.

Som GTS-institut nyder vi desuden en tillid i markedet, der potentielt muliggør, at vi kan få lov til at tilgå ressourcer hos danske virksomheder og frigive eksempelvis processeret/anonymiseret information og modeller baseret på disse ressourcer. Andre virksomheder nyder måske ikke samme uafhængighed og tillid. Et eksempel på dette kunne være virksomheder, der har unikke data fra deres kunder, så som regnskabsdata, og som kunne tænkes at blive forespurgt af IBM eller Microsoft om adgang til disse data med formålet at lave bedre automatiske sprogforståelsesressourcer. Denne virksomhed ville måske tøve med at give Microsoft eller IBM adgang, da dette potentielt også vil give Microsoft eller IBM adgang til forretningskritiske data omkring kunderne.

Relevante erfaringer og resultater

Alexandra Instituttet har en række relevante erfaringer og faglige ressourcer, der planlægges anvendt i dette projekt. Alexandra Instituttet har et stærkt fokus på data science/machine learning generelt via instituttets Data Science & Engineering (DSE) lab. En del af DSE-labs fokusområde er NLP-teknologi, og hvordan NLP-teknologi bringes bedst i spil til at hjælpe virksomheder med at bygge løsninger til automatisk sprogforståelse. Instituttet har i de seneste år haft særligt fokus på automatisk sprogforståelse og har blandt andet uddannet en ph.d. i deep learning på tekster samt ansat personer med speciale i NLP-

teknikker. I RK Big Data (2016-2018) har et af fokusområderne været automatisk sprogforståelse og NLP. DSE-lab har endvidere medvirket i [DABAI](#)-projektet, hvor vi bl.a. havde en projektlederrolle. Der har i DABAI-projektet været aktiviteter i at danne sig overblik over danske sprogressourcer, og der har været samarbejde med KU, AU og DTU, som der også planlægges samarbejde med i denne aktivitetsplan. Dansk for alle kan bl.a. ses som en forlængelse og adressering af de udfordringer, der er blevet afdækket i DABAI.

6) Nyhedsværdi og ambitionsniveau

Den nye teknologiske serviceydelse

Der er et stort samfundsmæssigt potentiale i at sikre viden og teknologi for danske virksomheder, som ønsker at arbejde med automatisk håndtering af det skrevne danske sprog. For at sikre danske virksomheders evne til at innovere og mulighed for at tilbyde deres kunder samme effektive service, som kendes fra udlandet, vil vi udvikle demoer/prototyper og modne teknologierne, så de nemt og hurtigt kan omsættes til løsninger for virksomhederne. Disse demoer/prototyper og teknologier vil i projektet tage udgangspunkt i konkrete behov hos virksomheder i særdeleshed de udvalgte casevirksomheder. Vi har identificeret følgende potentielle fokusområder:

- Question-Answering, automatiseret generering af svar og at finde facts i ustrukturerede eller semi-strukturerede dokumentsamlinger
- Dialogdrevne chatbots, bl.a. hvordan man holder styr på, hvor samtalen med brugeren chatbotten er kommet til.
- Semantisk søgning, hvad søger brugeren specifikt efter?
- Knowledge bases, population, søgning, vedligehold mv.
- Transfer learning, hvordan udnyttes viden fra et domæne i et andet domæne, potentielt hvor der er få høj kvalitets ressourcer til rådighed?
- Evidensbaseret NLP, løsninger som finder og kan tracke identiteter, entiteter, koncepter og argumentation over store dokumentbaser.

Vi planlægger at udbyde mindst 3 nye konkrete ydelser til gavn for både målgruppens virksomheder og rådgivere baseret på den viden og kompetencer, der opbygges i projektet. Forventede kategorier:

- Rådgivning af virksomheder omkring løsninger til automatisk håndtering af det skrevne danske sprog.
- Avanceret konfiguration og tilretning af NLP toolchains med specifik fokus på muligheder via vores åbne repositorium.
- Softwareudviklingsydelser omkring avancerede løsninger og komponenter til automatisk håndtering af det skrevne danske sprog. Specifikt komponenter, som ikke allerede er tilgængelige i det repositorium, der lanceres i projektet.

Foran markedet

Der er et stort behov for, at der udbydes ressourcer og viden omkring dansk NLP. Dette kan ses bl.a. ved, at en lang række SMV'er har udtrykt støtte til vores forslag på [bedreinnovation.dk](#) og givet udtryk for, at der er en mangel på danske sprogressourcer. En stor international virksomhed som IBM har også kommenteret. IBM ser et klart behov for en central opbygning af ressourcer og viden om automatisk

håndtering af dansksprogede tekster og ser, at det vil gavne alle virksomheder, hvis et sådant arbejde blev iværksat, og at store internationale virksomheder i mindre grad vil udvikle løsninger til dansk, hvis vi ikke iværksætter dette arbejde. Erhvervsstyrelsen har i august i år udsendt et notat, hvor de bl.a. beskriver udfordringen med NLP-ressourcer og påpeger, at adgang til sådanne ressourcer ”vil skabe stor værdi for danske myndigheder og virksomheder til udvikling af tidssvarende dansksprogede tjenester. Danske myndigheder vil således kunne levere bedre service til borgere og virksomheder”. Ressourcerne, der efterlyses, findes ikke på det danske marked i dag, og det er det behov, vi vil adressere i denne aktivitetsplan.

Tidshorisont

Det er instituttets klare forventning, at repositorium og serviceydelserne kan og skal markedsmodnes allerede inden for projektets 2-årige periode. Vi forventer at have basiskomponenter og ydelser udviklet og udbudt i projektperioden. Efter projektperioden vil der ske en yderlig udvikling og specialisering under almindelige kommercielle vilkår. Dette baserer vi på vores erfaring fra eksisterende resultatkontrakter samt den viden, vi allerede har opbygget i vores nuværende NLP-projekter bl.a. under EU/Region Midt projektet ”Innovativ brug af Big Data”.

7) Vidensamarbejde og -hjemtagning

Væsentligste samarbejder

Vi kommer i aktivitetsplanen til at arbejde tæt sammen med Københavns Universitet (KU). Derudover vil vi også samarbejde med Dansk Sprognævn (DSN) samt Center for Sprog Teknologi (CST) på KU. Som en del af aktivitetsplanen er det planlagt at medfinansiere en PostDoc eller lignende på KU.

DSN har en [arbejdsgruppe om moderne sprogteknologi](#) med deltagelse af en række danske virksomheder. Vi er allerede i dialog om at skabe synergi imellem deres arbejdsgruppe med tiltag i denne aktivitetsplan. Specielt er ATP og Erhvervsstyrelsen gennem deres deltagelse i arbejdsgruppen blevet til potentielle casevirksomheder i de planlagte aktiviteter i denne aktivitetsplan. DSN ser et behov og potentiale i Alexandra Instituttets casebaserede tilgang til at leverer input til arbejdsgruppen og den endelige rapport omkring de faktiske behov for ny sprogteknologi blandt de danske virksomheder. Denne aktivitetsplan kan således være med til at informere om fremtidige beslutninger på dette essentielle område til gavn for danske virksomheders konkurrenceevne.

CST har udtrykt, at de ser et potentiale i, at et GTS-institut som Alexandra Instituttet driver NLP-ressource repositorium, som måske i mindre grad ligger naturligt til den rene forskningsverden.

Vi får et meget tæt samarbejde med forskere med adgang til nyeste resultater og forskning. Vi planlægger også at inddrage bl.a. forskere fra KU til at informere og vælge tilgange til at løse caseprojekterne via workshops eller sparringsmøder.

Historisk har NLP-sprogteknologi og ressourcer ligget ved universiteterne, og der er et stort behov for at skabe overblik, nem adgang og en opdatering af tilgængelige ressourcer for at øge danske virksomheders kompetence inden for den næste generation af NLP-sprogteknologi. Dette er den primære grund til, at vi i denne aktivitet satser på et bredt, tæt samarbejde med universiteterne, også i højere grad end hvad der tidligere har været forsøgt. Vi ser universiteterne som en nøglepartner i det at hjælpe danske virksomheder med adgang til næste generation af dansk sprogteknologi.

Det er værd at bemærke, at vi oplever, at universiteterne også i høj grad er interesserede i dette samarbejde og ser dette samarbejde og denne formidling af deres ressourcer, datasæt, algoritmer mm. som en positiv bonus til gavn ikke kun for virksomhederne men også for universiteterne selv.

Samspil med andre projekter

Aktiviteterne bygger videre på viden opbygget i nuværende resultatkontrakter som bl.a. RK Teknologier og værktøj til udnyttelse af Big Data, hvor vi har arbejdet både med NLP og deep learning. Vi har arbejdet med nyeste deep learning metoder inden for rekursive neurale netværk, som finder særlig anvendelse inden for tekst, og vi har arbejdet med automatiseret håndtering af tekst i caseprojekter under projektet ”Innovativ brug af Big Data”. Vi bygger også videre på viden opsamlet i DABAI-projektet samt forventer fremadrettet synergi med Alexandra Instituttets center for anvendt kunstig intelligens (dansk.ai) og med fremtidige resultatkontrakter også.

Hjemtagning af viden

Vi har identificeret en række af de vigtigste internationale konferencer, hvorfra vi vil hjemtage ny viden i form af caseeksempler, industri- og forskningsresultater. Disse er ACL, EMNLP, NAACL, eller AAAI. Undervejs i aktivitetsperioden vil denne liste blive reevalueret bl.a. med input fra universitetspartnere.

8) Sammenhæng med instituttets strategi og afsæt i instituttets ressourcer

Denne aktivitet er central i forhold til Alexandra Instituttets overordnede strategi om at være danske virksomheders og forskeres foretrukne samarbejdspartner. Dansk for alle er også stærkt forankret i instituttets satsning på Big Data og kunstig intelligens og i instituttets Data Science & Engineering (DSE)-lab. I Alexandra Institutet ser vi NLP og sprogteknologi som et særdeles vigtigt område, og flere af vores udførende medarbejdere fokuserer primært på sprogteknologiske opgaver. Bl.a. har Alexandra finansieret opkvalificering af en af medarbejderne til ph.d. ved AU i deep learning på tekster.

Da DSE-lab kompetencemæssigt er klædt godt på til at takle de tekniske udfordringer, og da kunstig intelligens er et strategisk vigtigt satsningsområde for Alexandra Institutet med bevågenhed på direktionniveau, står Alexandra Institutet stærkt til at udføre denne aktivitetsplan og sikre danske virksomheder adgang til og viden om den næste generation af sprogteknologiske værktøjer. Derudover råder Alexandra Institutet over stærke antropologiske- og forretningsudviklingskompetencer, så også virksomhedsinvolveringen og udviklingen af nye ydelser vil blive varetaget på kompetent vis.

9) Tidsplan og milepæle

År 1

Vidensamarbejde, -hjemtag- og kompetenceopbygning

- Mindst 3 workshops eller sparringsmøder med forskere fra DIKU.
- Deltagelse i mindst én relevant international konference (f.eks. ACL, EMNLP, NAACL eller AAAI).
- Mindst 1 publikation indsendt ved internationale anerkendte, højkvalitets, peer-reviewed videnskabelige publikations venues (f.eks. ACL, EMNLP, NAACL eller AAAI).
- Kompetenceopbygning internt via teknologisk udvikling af komponenter til kommercielle ydelser, caseprojekter og repository. Mindst 1-2 komponenter udviklet. .
- Den nye viden fra disse aktiviteter dokumenteres i form af en eller flere populærvidenskabelige artikler, blogindlæg, indlæg i på hjemmeside for repository eller som videnartikler i Alexandras interne wiki (knowledge base) om NLP.

Udvikling af teknologisk service

- Online repository
- Mindst 10 ressourcer tilgængelige fra repository (dvs. datasæt, annoteringer, algoritmer, opskrifter, tutorials eller best practice beskrivelser).

- 2 nye kommercielle ydelser udbudt. Indhold vil justeres efter viden og markedsbehov. Vi forventer at kunne tilbyde ydelser inden for følgende overordnede emner:
 - Rådgivning af virksomheder omkring løsninger til automatisk håndtering af det skrevne danske sprog.
 - Avanceret konfiguration og tilretning af NLP toolchains specifikt med fokus på muligheder via vores åbne repositorium.
 - Softwareudviklingsydelser omkring avancerede løsninger og komponenter til automatisk håndtering af det skrevne danske sprog. Specifikt komponenter som ikke allerede er tilgængelige i det repositorium, der lanceres i projektet.
- Ydelser udbudt til mindst 3 konkrete leads.

Inddragelse og videnspredning

- Præsentation af ydelser og case/demo forløb online. Eventuelt både i nyhedsbrev og online på Alexandra Instituttets hjemmeside eller via hjemmeside for repositorium.
- Case/pilot-forløb med mindst 1 virksomheder
- Mindst 1 fagligt foredrag i relevant forum.
- Løbende seminarer, workshops, rundbordssamtaler eller studier med henblik på afdækning af virksomhedsbehov (minimum 2)
 - Heraf mindst ét større videnspredningsarrangement med minimum 20 deltagere
- Etablering af følgegruppe og afholdelse af 2 møder (f.eks. møde i Q1 og Q3).

År 2

Vidensamarbejde, -hjemtag- og kompetenceopbygning

- Mindst 2 workshops eller sparringsmøder med forskere fra DIKU.
- Deltagelse i mindst én relevant international konference (f.eks. ACL, EMNLP, NAACL eller AAAI).
- 1-2 publikationer indsendt ved internationale anerkendte, højkvalitets, peer-reviewed videnskabelige publikations venues (f.eks. ACL, EMNLP, NAACL eller AAAI).
- Kompetenceopbygning internt via teknologisk udvikling af komponenter til kommercielle ydelser, case projekter og repositorium. (Mindst 2-3 komponenter udviklet).
- Den nye viden fra disse aktiviteter dokumenteres i form af en eller flere populærvidenskabelige artikler, blogindlæg, indlæg i repositorium eller som videnartikler i instituttets interne wiki (knowledge base) om NLP.

Udvikling af teknologisk service

- Mindst 10 nye ressourcer tilgængelige fra repositorium (i alt minimum 20 tilgængelige).
- Mindst 1 ny kommerciel ydelse udbudt.
Ydelserne ligger ikke fast men vil tage udgangspunkt i erfaringer fra år 1. For eksempel kan der være tale om helt nye NLP-områder vi identificerer i samarbejde med danske virksomheder, eller der kan være tale om specialiseringer af de eksisterende ydelser eksempelvis til et nyt markedssegment.
- Ydelser udbudt til minimum 6 kunder.
- Mindst 2 demonstratorer klar til demonstration.

Inddragelse og videnspredning

- Præsentation af ydelser og case/demo forløb online. Evt. både i nyhedsbrev og online på Alexandras hjemmeside eller via hjemmeside for repositorium.

- Case/pilot-forløb. Mindst 2 nye i perioden.
- Mindst 1 fagligt foredrag i relevant forum.
- Løbende seminarer, workshops, rundbordssamtaler eller studier med henblik på afdækning af virksomhedsbehov (minimum 2-3).
 - Heraf mindst ét større vidensspredningsarrangement med mindst 20 deltagere.
- 1 møde med deltagere fra følgegruppen.