USA 9-19 november 2009

Sumithra Velupillai

Third i2b2 Shared-Task and Workshop

Challenges in Natural Language Processing for Clinical Data: Medication Extraction Challenge

13-14 november, San Francisco

Årets shared-task handlade om att extrahera läkemedelsinformation ur patientjournaler. Klasserna var: *medication*, *dosage*, *mode*, *frequency*, *duration* och *reason*. Nytt för i år var att annoteringarna gjordes som en *community annotation*, dvs alla deltagande grupper skulle bidra med annoteringar, som sedan användes i evalueringen, efter att ha skapat konsensus. Grupperna fick guidelines plus 17 exempelannoterade journaler. Under projektet fanns ett forum för diskussioner kring annoteringsarbetet. Av ca 40 registrerade grupper var det ca 20 som submittade resultat. Jon Patricks grupp från Sydney kom på första plats med resultat på 0.848 - 0.857 F-measure. Av grupperna på top 10 fanns representanter från tex Manchester och National Library of Medicine (NLM).

P. Zweigenbaum beskrev deras system som de kallade COKAINE, ett regelbaserat system byggt i Perl där de använde sig av externa lexikon för förkortningar, läkemedelsnamn, och tecken- och symptomlexikon från UMLS. De utgick från en 2-stegsmetod där de först identifierade läkemedel genom exakt match för att sedan hitta relaterad information (övriga klasser) mha reguljära uttryck och lexikon look-up.

Faisal beskrev ett system som baserades på CRF. De använde sig av särdrag som bigram på token-nivå, klassnivå, entitetstyp, frekvens. De använde sig av cTakes phrase parser (utvecklad på Mayo clinic). På grund avtekniska missöden kunde de inte vara med i evalueringen.

Ö. Nytröes grupp representerades av T. Brox Röst som bara varit delaktig i en liten del av arbetet (den som gjort det mesta kunde inte komma). Deras system var regelbaserat och byggde på olika delmoduler.

S. DuVall från VA (Veterans Affairs hälsodel) berättade om deras system som också var regelbaserat och utgick från, vad han kallade det, semantic bootstrapping, de letade efter frön och sen efter liknande mönster där dessa frön fanns. Systemet baserades på UIMA och cTakes, till vilka de lagt till några moduler. De filtrerade bort falska positiver och skapade topplistor av läkemedel.

I. Stolt berättade om systemet utvecklat vid Humboldt i Berlin då ingen av deltagarna i den gruppen kunde komma. Deras system var också regelbaserat, de skapade bland annat en egen grammatik för att klassificera entiteter och ytlig lingvistisk heurestik. För att identifiera negationer använde de det system de utvecklat för obesity challengen (2008).

S. Sohn beskrev systemet de utvecklat på Mayo, clinical nlp lab, som utgick från deras cTakes. Deras system var också huvudsakligen regelbaserat.

J. Patricks grupp från Sydney hade byggt ett kaskadsystem som använde både ML och regelbaserade metoder. De använde CRF för att identifiera entiteter, SVM för klassifikation av relationerna mellan läkemedlen och den relaterade informationen (de fem övriga klasserna), och en regelbaserad modul för att identifiera kontext. Han berättade att de i deras grupp har en NLP-infrastruktur med en modul för "knowledge discovery and reuse" där de utvecklat en metod för att identifiera okända enheter, tex ord. Han betonade vikten av att använda sig av lingvistiska strategier. Denna gruppen använde sig av Callisto (http://callisto.mitre.org/) för annoteringsarbetet.

A. Aronson från NLM berättade om deras system som var regelbaserat och där de använde sig av de resurser de utvecklat där, UMLS tex. För klassen reason använde de sig av MetaMap och Gopher.

H. Xu från Vanderbuilt beskrev systemet MedEx som de modifierat för denna challenge. I detta system görs preprocessning såsom meningstokenisering, semantisk klassificering och även regler för lexikonanvändning och en CFG parser. För att hantera felstavningar använde de sig av aspelalgoritmen.

Sammanfattning

På workshopen var det ca 30-40 personer, främst från grupperna som deltog men även några externa. Samtliga grupper hade väldigt låga resultat

på klasserna *reason* och *duration*. Dessa kräver mer sofistikerade metoder och det kom upp som förslag som fokus till nästa shared-task. Ambiguitet behöver kunna hanteras, så även anaphora resolution. Det var ganska livliga diskussioner kring huruvida ansatsen med *community annotation* var bra eller inte, det krävde ganska mycket jobb från alla deltagarna. Samtidigt spreds arbetsbördan och man kunde också ta fram gemensamm guidelines. Gruppen som kom på första plats (Sydney) var även de som var mest aktiva i annoteringarna, frågan om det eventuellt påverkade resultaten (bias) kom upp, svårt att analysera dock. Tanken med att inte ge deltagarna en helt färdigannoterad mängd var också att uppmuntra oövervakade eller semiövervakade metoder, men ingen av systemen byggde på sådant. Majoriteten av systemen var regelbaserade och utgick från diverse lexikon.

Jag kom direkt från Kanada till workshopen och missade de första talarna (några av dem presenterade dock på lördagen). De hade inga tryckta proceedings och man behövde användaruppgifter för att få tillgång till presentationer + publikationer (pga att undvika risk för att nekas publicering i journaler), Özlem skulle maila mig dessa men jag har inte fått dem än, har mailat och bett om dem och hoppas jag får dem när jag kommer tillbaka.

AMIA 15-18 november, San Francisco

Söndag, 15 november

Inledning och keynote

På årets AMIA var 1 900 deltagare registrerade (rekord!). Man kunde skicka in bidrag som student paper, poster, och paper. De hade samarbeten med flera journaler och en delmängd valdes ut som kandidater till dessa. De hade stort fokus på studentbidrag och delade ut pris för best student paper till tre studenter (från 80 bidrag skalades det ner först till 50 sen 8 sen 3). De har många olika priser, under inledningen delades Morris F. Collen award ut till Betsy L Humphreys från NLM.

Keynote gavs av dr. Mark Smith från California HealthCare Foundation. Hans presentation gick under temat *Hope, hype, how to avoid the road to hell*. Det var en mycket underhållande och inspirerande presentation där han pratade om vikten av detta forskningsområde, men också om att man måste vara realistisk och skapa saker som faktiskt är användbara. Trots bistra tider har just detta område ökat och framtiden spås vara ljus. Däremot är

detta området inne i ett skede av ”inflated expectations” och han tog upp tre viktiga saker:

- ”embrace affordability” - man måste komma ihåg att lösa det som går att lösa, och man måste vara medveten om att det är begränsade resurser man har att jobba med
- fokus konsumenten, dvs användarna: enkelhet, elegans
- få bort analogiteten, han gav exemplet med att ta tempen - som ju ofta görs med digitala verktyg - men sen skrivs det ner för hand...

S08: NLP and Data Mining

Ö. Uzuner presenterade pappret *Semantic Relations for Problem-Oriented Medical Records*, där de försökt modellera semantiska relationer på discharge summaries. De definierar relationer som handlar om patientens medicinska problem, representerade som *sjukdomar* och *symptom*. Klassificeraren bygger på SVM och utgår från en mening i taget. De använder ytliga lexikala och syntaktiska särdrag. De har kört systemet på journaler från både Beth Israel-Deaconess Medical Center (BIDMC), Boston, MA och Partners Healthcare, Boston, MA, och får lovande resultat inom spannet 0.68-0.96 f-measure.

S. Shon presenterade pappret *Mayo Clinic Smoking Status Classification System: Extensions and Improvements*, ett arbete där de utvecklat de system de byggde för 2006 års smoking status challenge. Detta system är baserat på cTakes. Klassificeringen gör för fem klasser: past smoker, current smoker, non-smoker samt unknown. Framför allt har de förbättrat negationsdetektionen för icke-rökare, identifikation av temporalitet, samt hanterande av klassen unknown. Systemet används för att identifiera riskfaktorer i en studie.

D. Hristovski presenterade pappret *Semantic Relations for Interpreting DNA Microarray Data*, ett system utveckalt som ett slags stöd för hypotesgenerering, eller åtminstonde för att stödja processen i att hitta relevant information för biomedicinsk forskning, här inom arbete med att tolka resultat från microarray-experiment, särskilt geninformation. Systemet bygger på discovery patterns, och de utnyttjar semantiska relationer (predications) hämtade från SemRep (skapat av NLM) ur MEDLINE-abstracts. Med det här systemet får användaren en djupare bild av hur gener relaterar till varandra genom de semantiska relationerna från SemRep.

C. Chute presenterade pappret *The Enterprise Data Trust at Mayo Clinic: A semantically integrated warehouse of biomedical data*, som beskriver Mayoklinikens Enterprise Data Trust (EDT), en stor datamängd som startade redan 1994. Det är en integrerad, top-down-byggd datasamling som används för analys och beslutsstöd på Mayo. De har olika delkomponenter, bland annat lexgrid som mappar standardterminologier mot det som finns i deras system. Chute fokuserade mycket på behovet av terminologier och processmodeller för att hantera stora mängder data inom hälsovården.

Måndag, 16 november

S20: Applications of NLP

M. Matheny presenterade pappret *Detection of Blood Culture Bacterial Contamination using Natural Language Processing*, ett arbete där de utvecklat ett system för att utvinna information ur microbiologi-journaler om blodkulturer, för att hitta antibiotika och bakterieinformation, samt bakteriella föroreningar. Systemet bygger på reguljära uttryck och Multi-Threaded Clinical Vocabulary Server (MCVS). De mappar bakterieinformationen mot koncept i SNOMED-CT. Systemet utvecklades iterativt, och resultaten för att mappa antibiotika och batkerieinformation var 84.8% (sensitivity), samt 96.0% (positive predictive value). För identifikation av bakteriell förorening fick de resultaten 83.3% (sensitivity) samt 81.8% (positive predictive value). De utnyttjade den hierarkiska strukturen i SNOMED och fick bättre resultat då de mappade mot en förälder i trädet.

J. Denny presenterade pappret *Development of a Natural Language Processing System to Identify Timing and Status of Colonoscopy Testing in Electronic Medical Records*. Trots att *colorectal cancer* är den andra vanligaste cancerformen, görs en *screening* bara i 40-60% av fallen. I det här arbetet utvecklar de ett nlp-system för att hitta fullföljda *colonoscopies* i journaler. De använde KnowledgeMap Concept Identifier för detta och systemet resulterade i 0.93 recall och 0.92 precision. Systemet preseterar bättre än om man utnyttjar faktureringskoder (billing codes) för att extrahera samma information.

H. Xu presenterade pappret *Extract Medication Information from Clinical Narratives*, samma system som använts i i2b2-tasken. Systemet utvecklades i första hand för i princip samma syfte som tasken, men systemet har mer detaljerad information kring läkemedelsinformationen, bland annat *dis-*

pense amount och *intake time*. I det här systemet har de använt NLTKs (http://www.nltk.org/) chart parser.

S. Agarwal presenterade pappret *FigSum: Automatically Generating Structured Text Summaries for Figures in Biomedical Literature*. Syftet med arbetet är att generera sammanfattningar av biomedicinsk litteratur för att få en bättre förståelse för vad bilder och figurer representerar, jämfört med om man endast ser bildtexten. Författarna har utvecklat en klassificerare baserad på multinomial naive bayes. Man utgår från att artikeln byggts upp enligt IMRaD (Introduction, Method, Results and Discussion), och systemet extraherar en mening från vardera kategori. Evaluerat på en manuellt annoterad testmängd fångar systemet 53% av meningarna samt resulterar i ett ROUGE-1-resultat på 0.70, högre än baseline. Systemet finns här: http://figuresearch.askhermes.org/

S30: Techniques for De-identification of Patient Data

J. Mayer presenterade pappret *Inductive Creation of an Annotation Schema and a Reference Standard for De-identification of VA Electronic Clinical Notes*, där de skapat ett schema för att gradera olika känslighetsgrad (hög-medel-låg) av PHI-klasser. De hade 9 olika typer av journaler, där vissa typer innehöll en mycket högre andel PHI (*discharge summaries*, *history* och *physicals*. 91.04% av annoteringarna klassades som medel, bara 4.46% klassades som hög. Bland dessa fanns patientnamn och *patient identifiers outside the facility*. De applicerade en induktiv och iterativ metod för att skapa detta annoteringsschema, där de reviderade efter olika tidsintervall. 2 annoterare, med ett snitt-IAA-resultat på 0.94 (mätt som 2*matches / (2*matches + non-matches)).

F. Morrison presenterade pappret *Using a pipeline to improve de-identification performance*. Här skapade de en pipeline, där de använde två olika existerande de-identifikationssystem för att automatiskt avidentifiera journaler; deid och MedLEE. Genom att använda båda system fick de bra resultat, bara två namn missades av systemet, dessa två namn var dessutom svårhanterliga (en initial och en ambiguös).

J. Liu presenterade pappret *Toward a Fully De-identified Biomedical Information Warehouse*, där skapandet av en avidentifierad information warehouse beskrevs. Här avidentifieras strukturerad information, och olika metoder testades: Secure Hash Algorithm (SH), Random Mapping (RM), Hashing-Mapping (HM) och Hashing-Mapping-and-Re-Mapping (HMM).

Det sista pappret i denna session, presenterat av C. Peng, *Assuring the Privacy and Security of Transmitting Sensitive Electronic Health Information* handlade om skapandet av en säker plattform för att kunna skicka hälsoinformation mellan olika institutioner. Plattformen kallas SRM (secure and reliable messaging platform) och bygger på OASIS, World Wide Web Consortium (W3C) web-services standarder, och Web Services Interoperability (WS-I) specifikationer.

S40: Semantic Modeling and Mapping

A. Aronson presenterade pappret *The Evolution of MetaMap, a Concept Search Program for Biomedical Text*, en slags statusrapport för MetaMap, utveklat på NLM. Med MetaMap kommer man åt koncept i UMLS-resursen Metathesaurus. Det finns en prolog-version och en Java-version, Prolog-versionen är mer extensiv, men långsammare. Systemet utvecklas kontinuerligt, och då Metathesaurus växer (från 44K till 2M!) uppdateras även MetaMap.

T. Cohen presenterade pappret *Predication-based Semantic Indexing: Permutation as a Means to Encode Predications in Semantic Space*, ett intressant arbete där de bygger en semantisk modell med semantiska predikat extraherade med SemRep (utvecklat på NLM, utnyttjar UMLS) mha Random Indexing, som de kallar PSI. Ur denna modell kan man sedan extrahera semantiskt relaterade koncept, tex för att hitta koncept som i modellen är nära associerad till ett specifikt predikat (tex ”treats”).

G. del Fiol presenterade pappret *A Large-scale Knowledge Management Method Based on the Analysis of the Use of Online Knowledge Resources.* När läkare vill kunna få mer information kring vad ett diagnostiskt testresultat betyder eller vilken information som finns kring detta, ska de kunna få det genom att trycka på en ”infobutton”, som länkar till externa källor. I det här arbetet hämtar de information ur Clineguide och ARUP Consult. 78.5% av fallen då en läkare tryckte på infobutton fanns information i någon av dessa kunskapskällor.

AMIA NLP working group, möte måndag kväll

Wendy Chapman från Pittsburgh är ordförande för denna grupp, Stephane Meystre är co-chair. I denna working group finns följande delgrupper: en ”Annotation group” som Özlem leder, plus två andra, en ”Application group”

som leds av Bob Futrelle och G. Savova, samt en ”Education and outreach group” som leds av S. DuVall.

Chapman började med att prata om planer som finns, bland annat har de sökt pengar för ett stort annoteringsprojekt där tanken är att det ska skapas en stor annoterad journalmängd som ska släppas för forskning. I Pittsburgh har de ett ”NLP repository”, blulab (http://www.dbmi.pitt.edu/blulab/). De har en avidentifierad mängd man kan få tillgång till, förutsatt att man själv har IRB approval (eller motsvarande) för sitt projekt. För tillgång till denna mängd måste man även bidra med eventuella ytterligare annoteringar man gör på mängden.

Tanken med annoteringsprojektet är:

- att skapa ett standardiserat schema, med guidelines, utvidga existerande standarder och försöka göra annoteringsprojekt interoperable
- att utveckla en manual för annoteringsmetodologi
- att bygga ett toolkit, inklusive evalueringsmått.

De ser gärna att man engagerar sig i detta arbete! För att definiera annoteringsklasser krävs en ontologi, eller någon form av modell, annars vet man inte vad man gör.

Verktyg för annotering finns, men de är många och olika bra på olika saker. Brian (minns tyvärr inte efternamn) har fixat med konsensusmodulen i Knowtator, den kan man få tillgång till. Man får gärna vara med i betatestning.

Özlem pratade om vilka uppgifter man egentligen vill att den här typen av datamängder ska kunna stödja. Hur ska man bäst utnyttja datat? Det är viktigt att dela med sig av sina problem och därmed undvika misstag. En katalog med scheman och guidelines som redan skapats för olika ändamål ska skapas.

I Application-gruppen ska en central resurs skapas där beskrivningar av hur nlp använts inom hälsodata ska finnas, vilken typ av data det finns, hur den använts osv. Ett slags ”repository”.

Det saknas standarder inom nlp för hälsodata - hur ska det appliceras? Definitioner av input/output-format, dokumenttyper, etc. behövs. Vilka PoS-taggar ska man ha? Tokenisering?

I Education and outreach-gruppen ska en bättre webbsajt skapas, och enklare socialt nätverk, förslag var att utnyttja LinkedIn. Det fanns även

försla på att ordna ”webinars” med inbjudna talare som jobbar inom området, som spelas in så att man kan komma åt dem olika tider.

S. Meystre pratade om att öka ”visibility” genom att sponsra papper, främja samarbeten med andra working groups inom AMIA, gemensamma projekt.

Förslag fanns även för att instifta ett award för best nlp paper, tex most innovative eller student nlp paper, fokusområden.

http://mailman.amia.org/mailman/listinfo/nlp-sig

Tisdag 16 november

S54: UMLS-NLP

H. Lowe presenterade pappret *Using a Statistical Natural Language Parser Augmented with the UMLS Specialist Lexicon to Assign SNOMED CT Codes to Anatomic Sites and Pathologic Diagnoses in Full Text Pathology Reports*, där de modifierat ChartIndex Medical Language Processing-systemet för att automatiskt identifiera anatomistka och diagnostiska substantivfraser ur patologirapporter och mappa dem till motsvarande SNOMED-koncept. För anatomiska koncept fick de resultat på for anatomic concepts 92.3% (positive predictive value), för diagnostiska koncept 84.4% (positive predictive value).

D. Denmer-Fushman presenterade pappret *UMLS Content View Appropriate for NLP Processing of the Biomedical Literature vs. Clinical Text*, ytterligare en slags progress report för de content views som finns för UMLS.

J. Geller presenterade pappret *Comparing Inconsistent Relationship Configurations Indicating UMLS Errors*, där de tagit fram en metod för att hitta inkonsekvent innehåll i UMLS-Metathesaurushierarkin, bland annat fel som att en förälder i trädet har en mindre generell typ än dess barn. Denna typen av problem kvantifierades och en hel del problem identifierades.

B. McInnes presenterade pappret *UMLS-interface and UMLS-similarity: Open-source Software for Measuring Paths and Semantic Similarity*, där de utvecklat ett system för att mäta semantisk likhet genom att mäta kortast möjliga väg (path) mellan olika koncept i UMLS-hierarkin. Systemet har implementerats i ett interface som kommer göras fritt tillgängligt.

S59: Applying Natural Language Processing in the Clinical Setting: An Overview

W. Chapman pratade om tillgängliga resurser och vikten av att ha bra annoterat material.

De-identification:

- i2b2 de-id challenge
- Pittsburgh deid (kommersiell)
- Mitre - Carafe
- Physionet
- miinlptoolkit (http://ostatic.com/miinlptoolkit)
- NLM

Shared task framtid:

- i2b2
- suicide notes (Cincinnatti, Pestian)
- VA
- AMIA NLP working group

Resurser:

- Cincinatti - computationalmedicine : http://www.computationalmedicine.org/project/nlp
- nlprepository Pittsburgh: http://www.dbmi.pitt.edu/blulab/nlprepository.html
- HiTEX (i2b2)
- cTakes MedKat (Mayo)

Carol Friedman höll G. Savovas presentation, hon pratade en del om framtida utmaningar:

- viktigt med gemensam representationsmodell - men där brister det ofta! kärnor såsom, sektion, mening, token, PoS, parsning.

- synonymer, varianter
- relationer, tid, co-reference
- kontext - osäkerhet, negation
- kliniskt relevanta händelser
- dok –> patient –> sjukdomsrelaterade metoder
- förenkla för läkare att skriva, då kan man använda nlp på ett annat sätt. Vad är en nlp-killer-app?

S75: Information Retrieval (två sista)

D. Kim presenterade pappret *Hierarchical Image Classification in the Bioscience Literature*, där de analyserat särdrag för automatisk klassificering av bilder i biomedicinsk litteratur. De har delat upp typer av bilder i: Gel-Image, Image-of-Thing, Graph, Model, och Mix. Först identifierar de textur för att klassificera bilderna i två grupper: texture (Gel Image, Image-of-Thing och Mix) och non-texture (Graph och Model). För första gruppen använder de sedan följande särdrag: entropy, skewness och uniformity. För den andra gruppen använder de edge difference, uniformity, och smoothness. Resultaten för den initiala klassificeringen i två grupper är lovande, den andra klassificeringsuppgiften gav inte lika bra resultat.

J. Fan presenterade pappret *Generating quality word sense disambiguation test sets*, där de utnyttjar MeSH för att annotera ämnesmässigt viktiga termer i biomedicinsk litteratur, i syftet att skapa testset för WSD. De rapporterar lovande resultat, särskilt om man cross-validerar med 2 eller 3 annoterare.

Onsdag, 18 nov

0.0.1 S86: Advances in NLP

S. Tu presenterade pappret *A Practical Method for Transforming Free-Text Eligibility Criteria into Computable Criteria*. I det här arbetet tittar de på fraser för att hitta lämpliga patienter som uppfyller lämplighetskriterier för specifika studier. De testar olika nlp-system för att automatiskt mappa

fritext till lämplighetskriterier enligt Eligibility Rule Grammar and Ontology (ERGO).

Harkema presenterade pappret *Methodology to Develop and Evaluate a Semantic Representation for NLP*, där de tagit fram en metod för att skapa en semantisk representation för att kunna extrahera relevant information ur journaler. Här jobbar de i tandvårdsdomänen, och skapar en ”information model”, som är en form av bottom-up-representationsmodell som skapas iterativt genom att manuellt extrahera de koncept som krävs för den specifika domänen, och sen iterativt utöka denna (och evaluera under tiden). De nådde ett snitt-IAA på 88%, och gjorde sammanlagt 12 ändringar i modellen och 20 ändringar i guidelines, bland annat för hantering av negationer.

R. Xu presenterade pappret *Unsupervised Method for Extracting Machine Understandable Medical Knowledge from a Large Free Text Collection*, en metod för att extrahera sjukdoms- och läkemedelsdefinitioner oövervakat, genom att utnyttja frasinformation. Metoden ”bootstrappar” genom att utnyttja ett seed pattern som sen förfinas genom olika rankningsmetoder.

R. Figueroa presenterade pappret *Tailoring Vocabularies for NLP in Sub-Domains: A Method to Detect Unused Word Sense*. Genom LSA tas semantisk närhet (neighborhood) fram i ett domänspecifikt vokabulär för tidigare okända ord. De lyckades identifiera okända ordbetydelser med en precision på 79% - 87% och recall mellan 48% och 74%,

S87: Latanya Sweeney - A Vision of Advanced Technologies for HIT

Sweeney pratade om framtiden för it inom hälsovården (för USA främst). Just nu finns det ingen koherent informationsstruktur i landet, utan ganska många isolerade öar. För tillfället satsas pengar på att försöka koordinera i alla fall delar av detta. Hon tog upp framtida behov som:

- medical billing framework
- personal health record
- centralized government database

Hon pratade även en del om privacy-frågor, särskilt om relationen mellan privacy OCH utility. Hon tror inte på ”ad-hoc”-de-identifikation (de-identifikation av existerande material), risken för re-identifikation är för stor.

Hon propagerade för att realisera de-identifikationen i de tekniska lösningarna, dvs. ”privacy by design: design markets and technologies with provable guarantees of privacy protection”. Enligt henne behövs nya typer av privacy-mekanismer och instrument. Hon pratade även om ”psychology of privacy”, det är ett känsligt ämne, patienten vill känna sig trygg i att ha kontroll, idén om privacy.

Särskilt intressanta posters

- Creating an Informatics Framework for Shared Decision-Making for Patient and Physician. Roxana Maffei and John H. Holmes
- Cognitive Study of Scientists' Use of a Knowledge Discovery Tool. Kavitha Mukund, Graciela Gonzalez and Trevor Cohen
- Cognitive Biases in Decision Making by Clinicians. Alexander Dragotoniu, David Robinson, Bhavesh Patel and Vimla L. Patel
- Annotation Schema for Anaphoric Relations in the Clinical Domain. Guergana Savova, Wendy Chapman, Jiaping Zheng, Melissa Castine and Rebecca Crowley

Studiebesök

Department of Medical Informatics & Clinical Epidemiology, Oregon Health & Science University, Portland 09 nov

I Portland höll jag ett seminarium för ca 6 personer: William (Bill) Hersh, Aaron Cohen, Steven Bedrick, plus några till. Min presentation utgick från det abstract jag mailat, och jag tog bilder dels från CEFAM, ISHIMR, licpresentationen. Jag berättade om Stockholm EPR corpus, om de-id-arbetet (med tillägg om LBM), de preliminära spekulationsresultaten (LREC 2010), hypotesgenereringsmetoden, och slutligen lite om Hexanord och Louhi/ACL. Kommentarer och frågor:

- Ang osäkerhet/säkerhet: se till att ta reda på *hur* läkare är säkra/osäkra - de har ett speciellt sätt att beskriva en del saker. Särskilt Hersh var mycket tveksam till detta och betonade att det är viktigt att ta reda på vad läkare faktiskt skriver (exempel: om en läkare skriver att patienten

förnekar något kan det betyda något helt annat). "Physicians learn specific ways of dictating".

- Perspektivfrågan väldigt viktig! Tydliga definitioner krävs - till vilket syfte vill man modellera (o)säkerhet? Vilken typ av (o)säkerhet är det som ska modelleras?
- Ang. hypotesgenereringsmetoden:
  - kan man lägga till någon form av relevance feedback?
  - hur göra för att få bort allt man redan vet? Risken är att det mesta man får fram med en sån här metod är sånt man redan känner till.
  - kan man koppla osäkerhet/säkerhet till detta?
  - andra representationsenheter än ord - vilka skulle det egentligen vara?
- Stort intresse för resurser, särskilt för cross-language challenge

School of Health Information Science, University of Victoria, Kanada, 12 november

I Victoria hade jag först ett möte med Denis Protti en stund. Han tipsade om WOHIT i Barcelona 15-18 mars: http://www.worldofhealthit.org/. HIMMS, EU, spanska regeringen. Höll sen ett seminarium (samma som i Portland, med några småändringar) för ca 10 pers, mest studenter. Kommentarer:

- Denis uttryckte stort tvivel till de-id och undrade varför det behövdes överhuvudtaget.
- Stort intresse för syntetiska journaler, framför allt för att kunna använda för att testa genom olika system.
- Synonymgrejen viktig, SNOMED eller andra typer av terminologier/ontologier kommer man nog inte undan.

Träffade även Andre Kushniruk som sysslar med väldigt intressanta saker. Usability-fokus, de har bland annat kunnat visa att en del medikationsfel berott på att användarna inte kunnat använda systemen ordentligt.

Institute for Health Informatics, University of Minnesota, Minneapolis, 19 november

I Minneapolis träffade jag Genevieve B. Melton-Meaux, hon är kirurg men också inblandad i NLP-projekt. Hon tyckte att (o)säkerhets-identifikation och -modellering verkade väldigt intressant men betonade också vikten av att definiera tydligt vad man vill modellera, bra om man kan samarbeta med någon kliniker, kanske bra också att börja med en specifik medicinsk verksamhet. Hon tyckte att det var viktigt att jobba mot någont form av terminologi. Serguei Pakhomov jobbar med många olika projekt för tillfället, dels semantiska relationer (mest via UMLS) och dels med akronym/förkortningsdisambiguering (i mindre skala för tillfället). Höll sen ett seminarium för ca 20 personer, mest studenter (inom hälsoinformatik). Lyckades få några att bli väldigt intresserade av att jobba med nlp! Fick tips om att det har gjorts en del jobb på namnhantering här (många svenska namn, svenska och nordiska immigranter), kolla med Dr. Martin Laventura på Minnesota dept. of state health, health informatics. Pakhomov var intresserad av innehållet i journalerna och undrade om det, i framtiden, skulle kunna gå att ordna tillgång till data även för folk overseas för specifika projekt, skulle nog finnas stort intresse inom AMIA för tex cross-language-studier.