tranScriptorium

Inleiding

Wereldwijd worden er enorme hoeveelheden handgeschreven historische documenten online gezet door digitale bibliotheken. Die onbewerkte digitale afbeeldingen worden pas écht nuttig als ze met informatie zijn verrijkt. Het project **tranScriptorium** richt zich op de ontwikkeling van innovatieve, rendabele oplossingen voor het indexeren, doorzoeken en volledig transcriberen van afbeeldingen van historische handgeschreven teksten. Het doet dit met behulp van Handwritten Text Recognition (HTR)-technologie. Voor een hoogwaardige transcriptie van handgeschreven teksten worden interactieve technieken ingezet die de gebruiker zo veel mogelijk werk uit handen nemen. De technologieën die hieruit voortkomen zullen worden geïntegreerd in een reeks HTR-tools. Die tools worden ingezet in een 'content provider portal' en ook in een gespecialiseerd HTR-webportaal voor transcriptie via crowdsourcing. Op die manier worden de optimale benutting en de validatie van de HTR-technologie gegarandeerd.

Voor handschriftherkenning zijn de bestaande Optical Character Recognition (OCR)-systemen niet geschikt. In afbeeldingen van handgeschreven tekst kunnen letters niet op een betrouwbare manier automatisch worden onderscheiden. Afbeeldingen van dit type vergen holistische benaderingen die vaak 'segmentation-free off-line HTR' worden genoemd, dus offline-handschriftherkenning waarbij niet wordt gesegmenteerd. Bij deze benaderingen moeten alle tekstelementen (zinnen, woorden en letters) als één geheel worden beschouwd, zonder ze vooraf uit elkaar te halen. De nieuwste HTR-technologie gebruikt concepten en methodes uit het veld van de Automatic Speech Recognition (ASR), zoals Hidden Markov Models en N-grams. Voor de toepassing van die technologie op handgeschreven tekstafbeeldingen zijn meerdere stappen nodig in de verwerking van die afbeeldingen. Gezien de enorme hoeveelheid data moeten die stappen zo veel mogelijk worden geautomatiseerd. Dat is een uitdaging, omdat er tot nu toe nog geen vaste regels zijn voor handgeschreven documenten. Zodra de tekstregels in een afbeelding gedetecteerd zijn, kunnen de HTR-modellen geheel automatisch worden afgeleid door middel van krachtige, veelgebruikte trainingtechnieken, waarvoor alleen de (volledige) transcripties van een relatief klein aantal tekstregels nodig zijn.

Doelstellingen

Ondanks recente grote verbeteringen is de huidige beschikbare HTR-technologie nog lang niet zo ver dat ze geheel automatische oplossingen voor transcriptie kan bieden. Om de nu nog ontoereikende HTR-technologie verder te ontwikkelen richt **tranScriptorium** zich op de volgende punten:

1. *Verbetering van de HTR-technologie t.b.v. efficiënte transcriptie.* Met de allernieuwste HTR-methodes als uitgangspunt zal **tranScriptorium** interactief-voorspellende technieken inzetten voor effectieve en gebruiksvriendelijke computerondersteunde transcriptie. Daarbij zal ook worden gekeken naar het automatisch of semiautomatisch genereren van gedeeltelijke transcripties van relevante tekstmetadata. We willen aantonen dat er goede resultaten kunnen worden geboekt in echte tekstverzamelingen zonder al te veel handmatige transcriptie en controlewerkzaamheden.
2. *Introductie van de HTR-technologie bij de gebruiker.* De verwachte gebruikers van de HTR-technologie zijn ruwweg in twee groepen te verdelen: a) individuele onderzoekers die ervaring hebben met de transcriptie van handgeschreven teksten en die specifieke documenten willen transcriberen. Deze gebruikers zullen bij het transcriberen veel baat hebben bij interactieve HTR-tools. Die tools worden als extra service naast de afbeeldingen ter beschikking gesteld op de site/webpagina's van de *content providers* (bibliotheken, archieven, e.d.). b) vrijwilligers die samenwerken aan grote transcriptieprojecten. Voor deze gebruikers worden de HTR-tools ter beschikking gesteld via een speciaal webportaal voor crowdsourcing dat een structuur biedt waarbinnen men gezamenlijk aan één project kan werken en experts de resultaten kunnen controleren.
3. *Integratie van de HTR-resultaten in openbare webportalen.* De HTR-technologie zal ondersteuning bieden bij het digitaliseren van het handgeschreven materiaal. De meeste digitale bibliotheken hangen tegenwoordig de resultaten van moderne OCR (*optical character recognition*) aan de gedigitaliseerde pagina's van tekstdocumenten; tot nu toe kunnen echter alleen afbeeldingen van *gedrukte* tekst op die manier worden gepresenteerd. In de toekomst kan dit ook met *handgeschreven* tekst, want de uitkomsten van de **tranScriptorium**-tools worden aan de afbeeldingen van gepubliceerde handgeschreven documenten gehangen. Daaronder vallen niet alleen volledige, correcte transcripties die tot stand zijn gekomen met de interactieve HTR-transcriptietechnieken, maar ook gedeeltelijke correcte transcripties en andersoortige automatisch gegenereerde metadata, die nuttig zijn voor het indexeren en zoeken gebaseerd op Key Word Spotting (KWS)-technieken. Deze (meta-)data zullen bezoekers en gebruikers in staat stellen op de klassieke manier in de volledige tekst te zoeken, te kopiëren en plakken, en tekst

af te drukken. Ze hebben dan het volle profijt van de beschikbare handgeschreven bronnen.

Binnen de beschikbare projecttijd willen we met **tranScriptorium** de ontwikkelde HTR-technologie toepassen op historische documenten in cursieve handschriften, waarvoor alleen HTR-technologie geschikte oplossingen kan bieden. De nadruk zal vooralsnog liggen op documenten waarvoor essentiële metadata of andere nuttige hulpmiddelen (deels) beschikbaar zijn: woordenboeken, (gedeeltelijke) transcripties uit andere bronnen, getranscribeerde tekst van dezelfde schrijver, enzovoort. **tranScriptorium** zal zich op vier talen richten: Spaans, Duits, Engels en Nederlands. We doen dat om de toepasbaarheid van de technologie op verschillende talen aan te tonen, maar ook om het gebruik en de validatie van de technologie te stimuleren bij een breder publiek.

Consortium

De partners binnen het **tranScriptorium**-consortium zijn zorgvuldig geselecteerd, zodat we voor alle verschillende taken de nodige expertise en middelen in huis hebben: onderzoekservaring op het gebied van HTR en DIA, content providers, transcriptie-ervaring en integratie-ervaring.

- Universitat Politècnica de València – UPV (Spanje)
- Universiy of Innsbruck – UIBK (Oostenrijk)
- National Center for Scientific Research "Demokritos" – NCSR (Griekenland)
- University College London – UCL (Groot-Brittannië)
- Institute for Dutch Lexicology – INL (Nederland)
- University London Computer Centre – ULCC (Groot-Brittannië)

Het INL en tranScriptorium

Het INL leidt een van de werkpakketten en is met name verantwoordelijk voor de lexica van de diverse talen en de taalmodellen die bij de handschriftherkenning gebruikt worden. Voor het Nederlands zal tranScriptorium zich met name richten op vijftiende-eeuwse arteshandschriften.

Projectduur

tranScriptorium is een STREP (*Small and medium scale focused project*) van het Seventh Framework Programme in de 'ICT for Learning and Access to Cultural Resources challenge'. Het project is gestart op 1 januari 2013 en zal duren tot 31 december 2015.

Website

www.transcriptorium.eu

Vacature

Research and Development Position Language Modelling for Text Recognition (pdf)