TRAINING NLP MODELS FOR THE ANALYSIS OF 16TH CENTURY LATIN AMERICAN HISTORICAL DOCUMENTS: TAGTOG AND THE GEOGRAPHIC REPORTS OF NEW SPAIN

P. Murrieta-Flores[1]; R. Liceras-Garrido[1]; M. Favila-Vázquez[2]; K. Bellamy[1]; J. Campos[3]; J.M. Cejuela[3]; B. Martins [4]

@patymurrieta; @RaquelLiceras; @Marianaarqueo; @kbellamy_; @Jorge_CP; @juamirocks; @bgmartins

1.-Digital humanities Hub – History Dep, Lancaster University (UK); 2.- Museo del Templo Mayor, Instituto Nacional de Antropología e Historia (MEX); 3.- tagtog.net; 4.- Instituto Superior Técnico and INESC-ID, Univ. de Lisboa (POR)

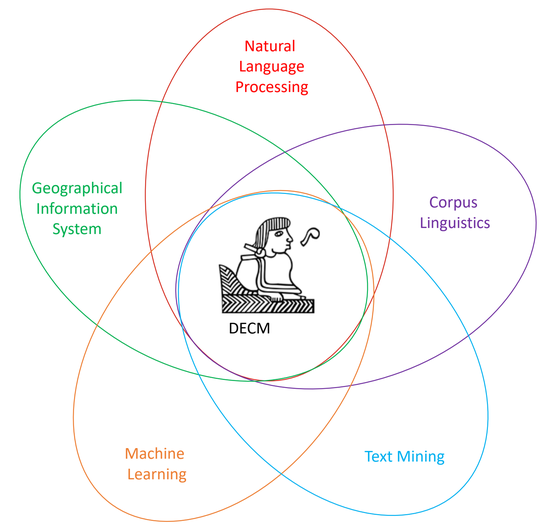
DIGGING INTO EARLY COLONIAL MEXICO (DECM)

DECM is a project that explores advanced computational techniques in conjunction with spatial analysis methods for the analysis of large historical corpora.

Using a Big Data approach, DECM focuses on analysing a 16th century corpus known as 'Las Relaciones Geográficas de la Nueva España' (hereafter RGs), consisting of documents related to New Spain, particularly the areas that encompass present-day Mexico and Guatemala.

Partnering with tagtog, we identify, extract and analyse textual information through a combination of techniques from Natural Language Processing (NLP), Machine Learning (ML) and Text Mining.

http://www.lancaster.ac.uk/digging-ecm/
@DiggingCH



AIMS

- Train NLP models to recognise the topics we are interested in analysing.
- Improve versatility when exploring text.
- Expand the scope of techniques like Named Entity Recognition (NER), nothing noting??? that these are typically only used to process modern newswire text and identify dates or proper names corresponding to places, institutions, and individuals.

CORPUS: THE GEOGRAPHIC REPORTS OF NEW SPAIN

The RGs include textual and pictorial information that portray the colonial situation of the domains ruled by the Spanish Crown, describing the life of their inhabitants and the state of the territories five decades after Spanish arrival.

This corpus contains around 2,800,000 words composed by the answers to the RGs questionnaire which were compiled between 1577 and 1585, covering topics including economy, resources, environment, traditions, geography, government, military organisation and language, among many others.

METHODS

The challenges of ML start with collecting training data. First, labelled datasets in this domain are scarce or non existent. Second, the increasing complexity and changing nature of linguistic nuances require the constant knowledge and verification from subject-matter experts (SMEs). In the context of NLP, this knowledge comes in the form of text annotations.

Main steps in the DECM approach:

1. Development of an ontology with a set of 22 tailored entities (information types) linked to dbpedia definitions and based on data relevant to our research questions.
2. Annotation of a data sample through tagtog, augmenting the historical documents with information on the spans of text corresponding to entities of interest.
3. Assess the inter-annotator agreement for the data sample, allowing us to spot biases and highly ambiguous cases, and assess also the occurrence frequency of the different classes (e.g. to check for unbalanced class usage or oversampled data)
4. Annotation of a second data sample, envisioning the subsequent training of NER models to automatically recognise entities and fine-grained labels.
5. Once the corpus is annotated and NER models are trained, we can proceed to the extraction of information for analysis.

Entity	Ontological definition	Labels
Person	http://dbpedia.org/page/Person	female/male, title, profession
Date	http://dbpedia.org/ontology/date	
Institution	http://dbpedia.org/page/Institution	civil, ecclesiastical, political, jurisdiction
Location	http://dbpedia.org/ontology/location	Settlement type, Generic location, geographic feature type, toponym, address, imaginary, jurisdiction
Activity	http://dbpedia.org/ontology/activity	agriculture, warfare, economy, mining, maintenance, female/male
Animal	http://dbpedia.org/ontology/animal	insect, mammal, reptile, bird, amphibian, aquatic, domesticated
Plant	http://dbpedia.org/page/Plant	
Food	http://dbpedia.org/page/Food	
Natural_Resource	http://dbpedia.org/page/Natural_resource	
Object	http://dbpedia.org/ontology/Thing	house_goods, commodity, clothing, weapon, tool
Architecture	http://dbpedia.org/page/Architecture	religious, civil, domestic
Cosmogony	http://dbpedia.org/page/Cosmogony	ritual, festivity, activity, deities, saints, object
Health	http://dbpedia.org/page/Health	disease, remedy
RouteOfTransportation	http://dbpedia.org/ontology/RouteOfTransportation	terrestrial, aquatic
Relationship	http://dbpedia.org/page/Relationship	spatial, kinship
Climate	http://dbpedia.org/page/Climate	
Ethnic_group	http://dbpedia.org/page/Ethnic_group	
Social_Class	http://dbpedia.org/page/Social_class	
Language	http://dbpedia.org/page/Language	
Event	http://dbpedia.org/page/Event	historical, disasters
RouteDirection	http://dbpedia.org/ontology/routeDirection	
Measurement	http://dbpedia.org/page/Measurement	value, tribute, weight, distance, population

FINAL REFLECTIONS

- We will soon have the full dataset of the RG's fully annotated and they will be made available as open data.
- Tagtog is a multi-user online tool with an easy-to-use interface, guaranteeing that scholars with low level of digital literacy are still able to participate and annotate documents.
- Including entities and labels in the annotation process enables to capture the complexities of the corpus and improves the accuracy of the results.
- After deciding on an annotation model, tagtog easily allows domain experts to address some challenges at the intersection of NLP, ML and CL.

tagtog http://www.tagtog.net @tagtog_net

Tagtog is a collaborative text annotation platform to find, create, and maintain NLP datasets efficiently. Accessible on the Cloud and On-Premises.

Collaborations between data analytics/AI professionals and subject-matter experts (SMEs) often fail. This is partially due to the lack of accessible tools, which could allow SMEs to participate in NLP tasks. To bridge this gap, tagtog was designed as a collaborative annotation platform with an easy-to-use interface.

Creating training data on tagtog is as simple as highlighting text. In addition, you can associate relations, attach attributes to entities, or classify whole documents. Annotations might be done both manually and automatically.

Automatic annotations reduce the effort required to produce labelled datasets. There are two methods available:

- **Dictionaries**: import or create collections of terms and extend them during the annotation tasks.
- **ML**: tagtog learns continuously from the provided annotations to generate precise predictions out of the box. If preferred, an external ML model can be plugged into the platform. SMEs review the ML predictions creating a continuous learning loop to train and keep the model up to date.

To quickly bootstrap annotation projects, tagtog supports several file formats natively. It enriches the annotating experience, eliminates unnecessary parsing steps, and allows users to annotate directly over PDFs, import PubMed articles, HTML, CSV, source code, or even Markdown files. For tighter integration, an API is available to import annotations and files, export annotations and metrics, and search.

To track annotation projects and data quality, tagtog measures the progress of the project members along with their agreement with other annotators (**Inter-Annotator Agreement**). Simply spot biases, unbalanced classes, or oversampled data by checking the distribution of your annotations.

CHALLENGES FOR NLP AND ML

- Our corpus is multilingual, written in 16th c. Spanish and peppered with words in 69 native languages. Techniques used to analyse large corpora have been developed and tested using modern English texts.
- RGs is a questionnaire, the topics discussed are multiple and diverse.
- Spelling variations. The same words are written in multiple forms.
- Semantic ambiguity. The same word can have different meanings, for instance the word '*vino*' depending on the context could be a conjugation of the verb 'to come' (he/she/it comes) or the noun 'wine'.
- The majority of words we are annotating are nouns. While the identification of proper nouns is well studied, the automated classification of words beyond these, that can be of use to specific text analysis, is more challenging.