A Method for Archaeological and Dendrochronological Concept Annotation using Domain Knowledge in Information Extraction

Andreas Vlachidis[a] and Douglas Tudhope[b]

[a] Department of Information Studies, University College London, Gower Street, London, WC1E 6BT, UK.
a.vlachidis@ucl.ac.uk

[b]School of Computing and Mathematical Sciences, University of South Wales, Trefforest, CF37 1DL, Wales, UK.
douglas.tudhope@southwales.ac.uk

Dr Andreas Vlachidis is a Lecturer/Assistant Professor in Information Science at the UCL Department of Information Studies. He has a long-term experience contributing to European and UK research projects focusing on cultural heritage data modelling and the multilingual application of Natural Language Processing in archaeological grey literature. His main research interests are in Information Extraction, Text Analytics, Knowledge-Based Systems and Ontologies. He is a certified text analyst, a fellow of the Higher Education Academy (FHEA) and a member of the British Computing Society (BCS). His research on the semantic indexing of cultural heritage resources has received several awards including the outstanding paper award from the Emerald Literati Network.

Professor Douglas Tudhope leads the Hypermedia Research Group at University of South Wales. He was PI on the AHRC funded STAR, STELLAR and SENESCHAL projects (Semantic Tools for Archaeological Resources) and led the Linking Archaeological Data Work Package for the ARIADNE FP7 Infrastructures Project. Since 1977, he has been Editor of the journal, New Review of Hypermedia and Multimedia. He co-authored the JISC State of the Art Review on Terminology Services and Technology and the JISC Terminology Registry Scoping Study. He was a member ISO TC46/SC9/SC8 (and NISO) working group developing a new thesaurus standard (ISO 25964). http://hypermedia.research.southwales.ac.uk/kos/

Abstract: Advances in Natural Language Processing allow the process of deriving information from large volumes of text to be automated. Attention is turned to one of the most important, but traditionally difficult to access resources in archaeology, commonly known as “grey literature”. This paper presents the development of two separate Named-Entity Recognition

(NER) pipelines aimed at the extraction of Archaeological and of Dendrochronological concepts in Dutch, respectively. The role of domain vocabulary is discussed for the development of a Knowledge Organization System (KOS)-driven, Rule-Based method of NER which makes complementary use of ontology, thesauri and domain vocabulary for information extraction and attribute assignment of semantic annotations. The NER task is challenged by a series of domain and language-oriented aspects and evaluated against a human-annotated Gold Standard. The results suggest the suitability of Rule-based KOS driven approaches for attaining the low-hanging fruits of NER, using a combination of quality vocabulary and rules.

Keywords: Information Extraction, Knowledge Organization Systems, Named Entity Recognition, Archaeology, Dendrochronology, Grey Literature, Semantic Annotation.

1. Introduction

Across Europe, the archaeological domain generates vast quantities of text in form of fieldwork and specialist reports often referred to as "grey literature" (Evans, 2015). Such literature is typically produced by government, academics and industry and published in print or electronic format for disseminating knowledge, not for profit, across a sector or a field of practice (Farace, 1997). For archaeological study such reports have significant advantages. They deliver in-depth analysis and discussion of excavation results without being restricted by page limits of conventional publications. However, access to the valuable information contained in such reports is a recognised problem. The detrimental effect on archaeological knowledge, as a result of the inaccessibility and difficulty in discovery of these texts, has begun to be increasingly recognised as a significant problem within the domain (Selhofer and Geser, 2014).

In the UK, the Online AccesS to the Index of archaeological investigationS (OASIS) project marked an early undertaking towards facilitating online dissemination via a unified repository (Richards and Hardman 2008). More recently, access to archaeological grey literature at the European level has been addressed by the aims of the ARIADNE Plus and its predecessor ARIADNE Infrastructure projects (Meghini *et al.*, 2017). The project has developed the ARIADNE Catalogue Data Model (ACDM) for providing an unambiguous representation of the archaeological information, integrating over 50K textual documents in its catalogue from a range of European countries. In the Netherlands, it is estimated that just under 60,000 of such reports have been produced over the last 20 years with a current estimated growth rate of 4,000 reports per year (RCE, 2017*)*. Searching over the corpus of Dutch Archaeological grey literature is being addressed by the AGNES system which employs Information Retrieval (IR) and Natural Language Processing Techniques for querying the documents with respect to specific archaeological concepts[1] and full-text matches (Brandsen *et al*., 2019).

This paper addresses the problem of facilitating access and information discovery in grey literature and implements a method for the automatic recognition of archaeological and dendrochronological concepts using vocabulary driven Natural Language Processing (NLP)

[1] The system provides faceted search for the following concepts; Artefact, Time Period, Location, Archaeological Context, Material and Species. http://agnessearch.nl

techniques. Such techniques have been recognised as vital for the automatic indexing, metadata generation, retrieval and dissemination from integrated online archaeological catalogues (Tudhope *et al*., 2011). Previous studies have employed similar methods for the semantic indexing of archaeological concepts in English with respect to domain ontology CIDOC-CRM (Vlachidis and Tudhope, 2016). Brandsen *et al*. (2019) explored the role of NLP for the development of effective search experiences in Dutch archaeological literature and agreed with our findings that such approaches can benefit indexing of archaeological grey literature for the purposes of retrieval and cross searching.

The distinct contribution of this paper is focused on the development of two separate NLP pipelines targeted at the recognition of archaeological and dendrochronological concepts in Dutch grey literature. The research is motivated by the aims of the ARIADNE Infrastructure project for making text-based resources more discoverable and delivers further contributions to the ARIADNE Plus project[2] by deploying the pipelines as cloud-based web services available to broader audiences. The method of the pipeline development is rule-based information extraction driven by domain vocabulary. An early version of the work (Vlachidis *et al.,*2021) was presented at the 14th MTSR conference (Garoufallou and Ovalle-Perandones, eds, 2021). The GATE (General Architecture for Text Engineering) framework (Cunningham *et al*., 2013) is used for configuring the NLP modules and to deliver interoperable semantic annotations which can be further analysed and used by consuming applications to identify patterns, trends, and "important" words or terms. Such interoperable outputs have been delivered by previous studies in English to facilitate archaeological information discovery, retrieval, comparison, analysis, and link texts to other types of data (Tudhope *et al*., 2011). The pipelines are available as standalone GATE applications from the ARIADNE services portal[3] and as cloud web services from the GATE Cloud[4].

The remainder of the paper is structured as follows. In Section 2, the background, main definitions and the role of vocabularies are discussed. In Section 3 the methodological choices are presented including adaptation of vocabularies to the NLP task, followed by a discussion on the structure of the pipelines. The paper concludes with evaluation of the archaeological concepts pipeline in Section 5, and a further discussion on the challenges, limitations and contributions of the paper in Section 6.

2. Background

The field of Information Extraction (IE) and particularly the task of Named Entity Recognition (NER) has been consistently growing over the past two decades. IE aims to identify instances of a particular prespecified class of entities, relationships and events in natural language texts, and the extraction of the relevant properties (arguments) of the identified entities, relationships or events. NER is specified as a subtask of IE - the term was first used during the Sixth Message Understanding Conference (MUC-6) (Grishman and Sundheim, 1996) to

[2] The EU ARIADNE Infrastructure project (EU FP7 Infrastructures, Grant agreement ID: 313193) was completed in 2013 and succeeded by the ARIADNE Plus project (EU H2020 Grant agreement ID: 823914) which will be completed in December 2022. https://ariadne-infrastructure.eu

[3] https://portal.ariadne-infrastructure.eu/services

[4] https://cloud.gate.ac.uk/shopfront#tagged=Ariadne%20Infrastructure

describe the task of extracting instances of entities of interest. Typically, the task has focused on recognition of instances of people, organizations, places, currencies, time and percentage expressions from unstructured text (Nadeau and Sekine, 2007). However, there is no single definition of NER as the task has kept on expanding and diversifying through the years to include additional entities such as products, events, diseases, to name but a few (Van Hooland *et al.*, 2015). In the context of archaeological fieldwork reports, the entities that have extracted the most interest relate to physical object, material, spatial and temporal information (Jeffrey *et al.*, 2009; Vlachidis and Tudhope 2016; Brandsen *et al.*, 2020).

Early NER systems employed hand-crafted rules supported by domain specific vocabulary resources for addressing the task of entity recognition (Nadeau and Sekine, 2007). Such vocabulary resources can range from simple flat-list glossaries to extended knowledge-based resources including thesauri and ontologies. Advances in Ontology-Based Information Extraction systems (OBIE) combine domain specialisation of instances with an inferential architecture of relationships to drive the task of entity extraction and to connect natural language to formal conceptual structures which are known as semantic annotations (Bontcheva *et al.*, 2004). The extraction engine of rule based systems is typically based on a pattern-matching mechanism for identifying and tagging entities with respect to knowledge-based resources, attaining good precision-focused performance at a relatively high system engineering cost (Piskorski and Yangarber, 2013).

Arguably, Rule-based NER systems may come with some scalability and domain adaptability limitations that potentially may be addressed with Machine Learning (ML) approaches. The typical supervised ML methods rely on training examples which are processed by a learning algorithm, for example a Hidden Markov Model or Naïve Bayes Classifier, to identify named entities in unstructured text. Therefore, the accuracy of the extraction is not dependent on the knowledge-base and rule performance but it heavily relies on the quality of the training corpus (Nadeau and Sekine, 2007). The labour-intensive process of designing a high-quality training corpus, which in many cases is subject to repeated human annotation and normalisation through inter-annotator agreement scores, has been addressed by ML approaches that require a small or no training corpus. These are known as semi-supervised and unsupervised respectively and are based on 'bootstrapping' techniques that start from a small training corpus or a seed, and expand their learning through iterations, exploiting contextual information for identifying entities that share similar characteristics and features (Toledo *et al.*, 2019). More recent examples of NER systems are built on the advances of Deep Learning methods, exploiting Recurrent Neural Networks architectures for factoring features of language models to address the task of NER (Fiorucci *et al.*, 2020).

In the broader field of humanities, NER has been employed to facilitate the study of historical corpora and to reveal social and spatial interconnections between people and places (Gelling, 2011). It has been understood as a method for discovering information over large data sets, across repositories and collections, via linking instances of common entities of interest (Angjeli *et al.*, 2014). In addition, the potential of NER for supporting the task of automatic generation of rich metadata has been recognised for disclosing information in large text collections, enabling semantic search of grey literature across disparate collections and datasets (Tudhope *et al.*, 2011). Such approaches can be employed to overcome time consuming metadata creation which may lack consistency when done by hand and when

created it is rarely integrated with the wider archaeological domain data. Moreover, the traditional model of manual cataloguing and indexing practices has been receiving less attention and priority. For example, prominent European research projects, such as the eContentplus explicitly did not fund the development of metadata schemas and the creation of metadata itself (Van Hooland *et al.*, 2015). Automated metadata generation via NER can compensate full text indexing techniques, enabling retrieval on multiple meanings and allowing researchers to search on concepts taking account for synonymy and polysemy (Tudhope *et al.*, 2011; Brandsen *et al.*, 2019; Jeffrey *et al.*, 2019).

A number of projects have employed IE and NER techniques on archaeological grey literature. An early pilot application was carried out by Amrani *et al.* (2008) that used an IE pipeline composed of part-of-speech tagging, co-reference resolution, terminology extraction, classification of concepts and validation for extracting archaeological concepts of construction, period, site, method and solid type. The OpenBoek project experimented with memory based learning to extract chronological and geographical terms from Dutch archaeological texts (Paijmans and Wubben, 2007). Byrne and Klein (2010) also investigated the extraction of information from archaeological literature primarily focusing on extraction of events from unstructured text. The Archaeotools project adopted a machine learning approach to enable access to archaeological grey literature via a faceted classification scheme of What (what subject does the record refer to), Where (where, location, region of interest), When (archaeological date of interest) and Media (form of the record) which combined databases with information extracted from reports in an interesting faceted browser interface (Jeffrey *et al.*, 2009). The OPTIMA system applied a Rule-based, Knowledge Organization System (KOS) driven approach to semantic indexing of English language archaeological grey literature, using named entity recognition, relation extraction, negation detection and word-sense disambiguation techniques (Vlachidis, 2012). The system employed the semantics of the CIDOC Conceptual Reference Model (CRM), a standard (ISO 21127:2006) ontology for cultural heritage, for associating ontology classes to textual instances which were further complemented by terminological and typological definitions based on English Heritage thesauri and domain glossaries. Moreover, the Information Retrieval system AGNES has employed NER techniques for complimenting full-text indexing of Dutch archaeological grey literature to enable semantic retrieval with respect to artifact, time period and material entities (Brandsen *et al.*, 2019).

The NER pipelines discussed in this paper develop from the OPTIMA system and expand the Rule-based KOS driven approach in the context of Dutch archaeological and dendrochronological entities. The distinct contribution of the new pipelines compared to previous work reside on their standalone and system independent operation. The pipelines are not coupled into a particular architecture, system or project and are available as configurable, adaptable and expandable open source GATE applications and as cloud web-services, allowing for versatile use within Dutch archaeological grey literature. In addition, to the best of our knowledge, the Dendrochronology NER is one of the first pipelines targeted at recognising material and sample entities relating to this particular archaeometric technique. Dendrochronology is a method for date detection based on tree ring identification which is highly quantitative as a discipline and has been supported by computational methods to perform tree-ring measurements, functions and statistical analyses (Kuniholm, 2002; Bunn,

2008, Jansma et al., 2012). The NER pipeline is focused on extracting concepts following the results of such analyses that are reported in grey-literature documents.

The Dutch pipelines are challenged by language features that directly affected recognition of compound noun forms, place names and time entities. Although, a significant amount of effort has been spent on information extraction in English, covering NER and higher-level IE tasks, such as relation and event extraction, arguably less attention has been paid to non-English languages (Piskorski and Yangarber, 2013). The performance of non-English IE systems is challenged by linguistic phenomena, such as productive compounding, which complicates morphological analysis in German and Dutch and proper name declension forms in Greek and Slavic languages which complicate named entity recognition (Piskorski *et al.*, 2009). The following sections discuss the methods and techniques used to address some of these challenges in order to extract entities of interest from Dutch archaeological grey literature.

3. Methodology

The Information Extraction method of the NER pipelines is described as Rule-based KOS driven. It is a method that shares similarities with the Ontology Based Information Extraction (OBIE) approach which uses an ontology as the basis for domain knowledge. OBIE links the extracted instances to ontological classes, enabling meaningful representations that can be used as relational information or to perform reasoning (Bontcheva *et al.,* 2004). Our approach extends OBIE by making a complementary use of ontology, thesauri and domain vocabulary for driving the NER task and standardising the annotation of the extracted textual instances with references to domain terminological resources. The rule based approach utilises the Java Annotation Pattern Engine (JAPE), which is a finite state transducer native to the GATE framework we used for the construction of the pipelines. JAPE grammars contain pattern-matching expression in form or regular expression for detecting textual snippet. They contain a LHS (Left Hand Side) for handling the regular expressions and a RHS (Right Hand Side) for manipulating the annotations matched by the rules which exploit lexical, syntactical, morphological, vocabulary and ontological features in rich pattern-matching expressions.

The employment of a rule-based approach driven by domain resources distinguishes our approach from supervised machine learning methods, which heavily rely on the existence and quality of training data. The definition of training sets for supervised ML requires specialised human input and management of the tagging process to ensure the quality of training. A training set for Dutch archaeological grey literature was developed by Brandsen *et al.* (2020) employing archaeology students for tagging a corpus of 31,000 instances for the entities; Artefact, Time Period, Location, Context, Material, and Species. At the time of development of the Archaeological and Dendrochronological NER pipelines such a training corpus was not available. Our design choice to follow a rule-based approach was influenced by the availability of quality domain vocabulary resources and the lack of a fully developed ML training corpus. Our experience demonstrated the suitability of rule-based, vocabulary driven approaches in the domain of archaeological NER for English, delivering operational performance for a range of entities with particular focus on precision and extraction of rich instances carrying quantifiers and moderation elements of an instance (e.g. *large pit, roughly around 200BC,* etc.) (Vlachidis and Tudhope, 2012). Although, development of rule-based systems requires

investment of specialised computational skills, this requirement is also pertinent to ML approaches. A major advantage of vocabulary driven rule-based systems is their focus on domain precision of entity recognition. The performance of such systems is influenced by the coverage of the domain vocabulary which can impact recall for frequently occurring terms in text that are not available in the vocabulary. Nonetheless, their portability to new tasks within the domain is fairly straightforward. The lexical and morphological behaviours targeted by the pattern-matching rules usually remain applicable or can be adapted whilst vocabulary resources can be added and removed as "cartridges" to drive a new task.

The Archaeological-concepts NER pipeline is designed to extract core entities of research interest within the context of Dutch grey literature. The list of entity types includes; Artifacts, such as finds and objects (e.g. pottery/*aardewerk*, bone/*bot*), Archaeological context (e.g. posthole/*kuilen*, ditch/*sloot*), Materials (e.g. charcoal/*houtskool*, iron/*ijzeren*), Monuments types in the sense of physical structures (e.g. church/ kerk, settlement/*nederzetting*), Places in the sense of geographical locations (e.g. Leiden, Bunschoten), and Time Periods (e.g. Stone Age/*Steentijd*, 5300 - 2000 vC). The Dendrochronological-concepts pipeline specialised the NER approach to extract entities of timber including both wood material of tree types (e.g. oak/*eik*, beech/*beuken*), wood products and architectural elements (e.g. lumber/*timmerhout*, pole/*paal*), numeric date values such as '1020 AD' and mentions of wood sampling (e.g. *dendro-monster*). In addition, the pipeline extracts phrases containing mentions of two or three different entities types of interest such as Material, Architectural Element and Date. As for example in the sentence, 'oak from which the pole is made was felled between 55 and 69 AD' (*eik waaruit de paal is vervaardigd, is geveld tussen 55 en 69 na Chr*). Extraction of such phrases was based on the co-occurrence of entities within the context of a single sentence without employing syntactical patterns. The NER task is focused on extracting entities from specific document sections containing relevant discussion of dendrochronological analysis. This particular approach is followed for improving the precision of the task and to eliminate matches from areas less relevant to dendrochronology which are contained in long post-excavation documents that report on a range of phases and workflows. The specific details of the pipelines and their cascading arrangement of modules and NLP processes is discussed in Section 4.

The Archeologisch Basis Register (ABR) thesaurus is employed as the main knowledge-based resource for driving the NER process of Archaeological concepts (Erfgoedthesaurus, online). The thesaurus is maintained by the Dutch State Service for Heritage (Rijksdienst voor het Cultureel Erfgoed - RCE) and contains 20,000 terms (excluding synonyms) divided into seven main areas (facets) that serve a frame of reference for heritage concepts, including abstract concepts, activities, physical characteristics, materials, objects, actors and periods. The thesaurus is available as a searchable web-based service and as Linked Data resource from the RCE servers[5], and as a Simple Knowledge Organization System (SKOS) resource from the OpenSKOS project[6]. Several facets and subareas of the thesaurus have been selected and used by hand-crafted rules to drive recognition of the targeted entities, including Artefact Types and Context (*Artefacten*), Monuments (*Complextypen*), Periods (*Perioden*), and Materials (Materialen).

[5] https://thesaurus.cultureelerfgoed.nl/search;schemes=abr:b6df7840-67bf-48bd-aa56-7ee39435d2ed
[6] http://openskos.org/api/collections/rce:EGT.html

The Getty Arts and Architecture (AAT) thesaurus[7] is employed by the Dendrochronology NER pipeline for driving the process with concepts of wood products, wood materials and architectural elements, using the corresponding thesaurus facets. The AAT thesaurus is a standard structured vocabulary of generic concepts related to the domains of arts and architecture, including archaeology, conservation and the broader cultural heritage (AAT, online). The AAT concepts enjoy multilingual labels, including Dutch, which were used for creating the vocabulary resource of the NER pipeline. An additional set of vocabulary resources was also made available and used as part of the collaboration activity within the ARIADNE Infrastructure project, including a gazetteer of Dutch placenames and a list of supplementary Dendrochronology related concepts. Pipeline-specific glossary resources also have been constructed containing period related suffixes to support recognition of temporal instances and numerical dates e.g. '1200 AD', '800 v.Chr', etc.

Employment of the thesauri enables assignment of a range of attributes on each individual recognised concept, including information about the origin of a term (contributing thesaurus), unique reference (URI) and a corresponding terminological reference to the thesaurus which is uniquely identified by URI. The output of entities is delivered in a structured and interoperable XML format, constituting a document index. Such output can be used in information retrieval and further analysis of the grey literature documents (Tudhope *et al.*, 2011) or consumed as training set by ML applications (Brandsen *et al.*, 2019).

3.1 Adapting KOS to NLP

Importing the ABR thesaurus into the GATE framework required adaptation of the resource into the NLP task both for optimising matching of concepts and for enabling use of the thesaurus relationships by JAPE rules. The process involved transformation (serialisation) of the thesaurus structure to Ontology Web Language (OWL-Lite) format and adaption and enrichment of the thesaurus labels to the natural language context as discussed below.

Serialisation of the thesaurus to OWL-Lite was required for enabling the rules to exploit the broader/narrower semantic relationships of the structure. The original RDF serialisation of the thesauri can be only partially parsed from the GATE framework, causing the rich thesaurus structure to flatten, preventing rules from making use of the thesaurus semantics. Expressing the structure to OWL-Lite serialisation not only enables use of the hierarchical relationships but also facilitates matching on alternative labels, synonyms, and assignment of interoperable attributes already available in the original resources, such as SKOS unique identifiers. The process also created new human-readable uniform resource identifiers (URIs) while maintaining the original references for individual entries (i.e. *rna:contentItem* and *skos:Concept* and *rdf:about*). The necessity to provide new human-readable URIs for classes is dictated by GATE's behaviour towards exposing class URI to JAPE rules.

The transformation process employed XSL templates for creating the new OWL-Lite serialisation of the thesauri resources. The process assigned human readable URIs to the ontology instances and mapped the thesauri SKOS semantics to standard RDFs terms. The process created the human readable URIs by combining a temporary base URI with the

[7] https://www.getty.edu/research/tools/vocabularies/aat/

preferred label of thesauri terms that previously have been cleaned of illegal characters (e.g., ampersand, slash, space etc). The *skos:Concept* property has been mapped to the *dcterms:identifier,* for holding the unique terminological reference of a thesaurus entry and the *rdfs:seeAlso* annotation property is used for holding any additional references of the term. In addition, the *rdfs:subClassOf* property has been used for implementing the hierarchical relationships of the resource.

Adaptation of the ABR thesaurus to the requirements of the NLP was required due to the construction of the ABR terms (SKOS labels) which follows a classification descriptive approach rather than using labels closer to natural language. The thesaurus was not developed with Natural Language Processing in mind and as a result contains labels that are not suitable for automatic and algorithmic term matching due to their multiterm, sometimes descriptive and verbose punctuation structure. The adaptation effort was focused on resolving such overloaded vocabulary entries into individual term components. For example, the vocabulary entries 'amulet/talisman' and its child entry 'amulet/talisman – kruisvormig' (cruciform) do not correspond to the way in which such terms are used in natural language text. Most likely either amulet or talisman will be found as individual entries and if an adjective is used, such as 'kruisvormig' this will follow a grammatically correct syntax form (i.e. 'kruisvormig amulet' instead of 'amulet kruisvormig'). Vocabulary entries like the above were enhanced with labels that are closer to what is likely to appear in text rather than carrying descriptive and non-natural language descriptions.

The process of adaptation and label-enrichment employed Extensible Stylesheet Language Transformations (XSLT) aimed at label patterns that joined synonyms and specialisations together under a single label. For example, the forward slash (/) character joins synonyms as in the case 'amulet/talisman', the hyphen (-) character adds specialisation as in the case 'amulet/talisman – kruisvormig' and the comma (,) character adds a form of periphrastic description which can be treated as an alternative label. The XSL templates incorporated the above patterns to generate the new vocabulary labels where for example 'amulet/talisman' delivers two separate labels (i.e. amulet, talisman) and 'amulet/talisman – kruisvormig' delivers the labels 'kruisvormig amulet' and 'kruisvormig talisman'. In most cases, special characters for joining synonyms and expressing specialisations or generalisations are consistently used across the ABR thesaurus and the transformation delivered useful alternative labels. However, there are cases that do not follow the standard use of special characters or are very verbose (e.g. 'hu-isplattegrond:4-schepig - type St.Oedenrode'). Such cases due to their complexity were not matched by the transformation templates and were ignored.

4. The NER Pipelines

The GATE framework provided the development environment for building the pipelines, making available the necessary tools and NLP modules for arranging the NER task in a cascading order of subsequent processes. At the core of the process resides an ontological structure which accommodates the vocabulary and semantic relationships for driving the NER task. JAPE grammars can invoke the hierarchical relationships of the ontology transitively for creating matches that conform to the semantics of a particular ontology class and its sub-classes. Hence, a syntactically simple matching expression can be very versatile to create

matches from a hierarchical branch. For example, a rule can be focused on a specific monument type e.g., “Defensive Structures” in order to match vocabulary instances that correspond to the class and its sub-classes, such as “castle”, “tower”, etc., whilst ignoring any other matches beyond the focused class and sub-classes. This particular agility of rules allows for mapping between ontology classes and named-entities, enabling precision and flexibility for consuming only the relevant vocabulary for an entity using a small number of rules. In addition, individual ontological classes or instances benefit from the use of parameters holding spelling variations, synonyms, SKOS identifiers and any other bespoke parameters useful to the NER task.

A range of general purpose, domain independent NLP modules are employed by the pipelines for creating a set of features that are utilised at subsequent stages by entity-focused rules. Such domain independent modules are situated at the early stage of the pipeline (Pre-processing) for identifying sentence boundaries, tokenise text into individual words, annotate words with part-of-speech categories and morphological features such as lexical stems. The pipelines run in a cascading order where each module adds a layer of semantics to the output and the order of the modules is important. For example, the lexical stem output is critical for the operation of the ontology module for delivering lookup annotations with respect to targeted classes. The ontology lookup output is then available to the named-entity rules which combine lookup and other token input for precision and performance. The details of the two NER pipelines are discussed in the following section.

4.1 Archaeological Concepts NER

The archaeological concepts NER pipeline is made of 3 main phases namely; Pre-processing, Lookup matching, Named-entity extraction and validation. The pre-processing phase incorporates a word tokenizer, a sentence splitter, a part-of-speech tagger and a stemmer, in that order. With the exception of the stemming module, which is based on the Snowball[8] stemmer for Dutch (Kraaij and Pohlmann, 1994), the modules are based on the Apache Open NLP[9] toolbox. All modules are wrapped for the GATE framework and are available from the respective plugins directory. Upon completion of the pre-processing phase, individual words are annotated with the following set of features; *category* that carries the part of speech tag, and *length*, *stem* and *string* of a word. For example, the word ‘canals/*grachten*’ is assigned the features *category*: N (i.e. noun), *length*: 8, *stem:* gracht and *string*: grachten.

The next phase of the pipeline is focused on creating the Lookup matches that originate from the ontology and the supplementary vocabulary resources (i.e. GATE Gazetteers). During this phase the matches are not aligned (linked) to a specific name-entity but are rather high-level matches of vocabulary lookup that contain features reflecting their thesaurus origin and ontology class. For example the Lookup for the word ‘*grachten*’ is assigned the features; *Class*: http://tmp/Artefacttypen#gracht of the internal URI of the ontology, *Identifier*: https://data.cultureelerfgoed.nl/term/id/abr/78b755c1-9f32-42b9-b243-1bf6764af484.html of the external ABR thesaurus reference, and *Thesaurus*: complextypen ABR reflecting the origin of thesaurus facet. In addition, the supplementary, pipeline-specific

[8] https://snowballstem.org/
[9] https://opennlp.apache.org/

glossary is invoked in the phase for producing lookups of temporal suffixes such as 'AD', 'BC', 'c14', etc. for aiding recognition of Time Period entities.

The final phase of the pipeline is targeted at extracting the named-entities of interest using dedicated rules for each entity type. The first stage of the phase invokes rules that hard-code a stop-list of terms which are excluded from matching. The stop-list contains terms that during development have been identified to affect precision due to their non-specific focus. Such terms appear as alternative labels of ontology classes and include 'A/*Een*', 'weight/*gewicht*', 'can/*kan*', and other words of similar generic scope. A cascading order of JAPE rules is then invoked to extract the entities utilising stop-list, part of speech category and targeted instances from the ontology. The rule responsible for the extraction of Archaeological context uses instances from a range of ontology classes, including 'canal/*gracht*', 'ditch/*greppel sloot*', 'pit/*kuil*', 'pole wreath/*paalkrans*', 'stockade/*palisade*' and other relevant to context classes. The left-hand-side of the rule holds the actual structure of the matching pattern which targets instances in text that are not in the stop-list (i.e. NotLookup type), are nouns and are instances originating from a specific class and subclasses of the ontology (transitive matching).
For example `{!NotLookup, Token.category == N, Lookup.class == [rceFeatures n= gracht]}.`

The right-hand-side of the rule is responsible for passing to the named-entity annotation the lookup features *URI*, *Identifier* and *Thesaurus* and to assign the respective entity type, for example Archaeological Context. For some entity types the left-hand-side of the rules exploits transitively the structure of broad-level classes, as for example in the case of Material, where every instance of the Material facet participates in the pattern-matching rules. Other types, such as the numerical date of the Period type invoke several complex pattern-matching rules to extract variation of instances as seen by the rule below which addresses textual instances that contain numbers and data suffixes (e.g. '5300 - 2000 vC', '800 v.Chr. - 1500 n.Chr').

```
{Token.category==Num}
({Lookup.majorType == periodSuffix})?
({Token.string == "-"}|{Token.string == "/"}|{Token.string ==
"–"})?
{Token.category == Num}
{Lookup.majorType == periodSuffix}
```

4.2 Dendrochronological Concepts NER

The Dendrochronological NER pipeline follows a similar cascading arrangement to the Archaeological Concepts pipeline and is split into the three main phases. The pre-processing phase uses the exact same NLP modules; Tokenizer, Sentence Splitter, Part-of-Speech tagger and Stemmer. The Lookup phase utilises ontology classes originating based on the AAT thesaurus instead of the ABR thesaurus. In addition, during Lookup specific document sections appearing to contain rich dendrochronology related discussion are identified using heuristic rules. The rules use a list of relevant concepts, such as 'dendronchronology analysis/ *dendrochronologische analyse*', 'analysis dendrochronological/ *analysendendrokronologis*',

'*dendroproverna*', 'cut date/*kapdatum*', 'update method/*updatemethode*' and other similar 38 in total terms. The identified section expands 3 sentences above and 3 sentences below the dendrochronology relevant concept, as seen in Figure 2.

The named-entity phase is focused on extracting concepts of wood material and products (Material), architectural elements (ArcElement), numerical instances of dates (Date) and mentions of sampling activity (Sample). The rules run in a cascading order and follow the same pattern matching approach with the Archaeological concepts NER previously discussed, focusing on instances that are non-stoplisted, are nouns and originate from targeted ontology classes. In addition, the phase uses a set of rules to identify phrases that contain a combination of two or three different entity instances of the types Material, ArcElement and Date in any order. Such phrases are given a weight 100 and 60 for containing 3 or 2 entities types respectively (Figure 2).

5. Evaluation

The performance of the Archaeological concepts pipeline was benchmarked via a Gold Standard (GS) set of manual annotations defined for the purposes of the ARIADNE project by a group of Dutch archaeologists (Leiden University). The Gold Standard refers to a set of human annotated documents which represents the desirable result and is used for comparison with system produced automatic annotations. It consists of 7 long grey literature reports containing approximately 4,000 annotated instances of several entity types, including Archaeological Context (Feature), Artefact, Event, Material, Method, Monument, Place, Period and Person. The entities Event, Method, and Person were not in the scope of the NER pipeline and are not included in the evaluation. Precision, Recall and F-measure (weighted average) are used for reporting the system's performance. The metrics are well-established and originally introduced by the second Machine Understanding Conference, MUC 2 for measuring the performance of information extraction systems (Grishman and Sundheim, 1996). The scores are calculated as fractions of *True Positive (TP)*, *False Positive (FP)* and *False Negative(FN)* matches. A TP indicates a correctly identified entity match whereas a FP indicates an erroneous entity match (i.e. match that does not correspond to the correct entity type - mismatch). A FN indicates that an entity is available in text but not identified by the system (i.e., a total miss - not a mismatch). TNs are all mentions not relevant to the scope of the NER task (i.e. all entity types or otherwise that are not addressed by the system) and as such do not contribute to the scores and are not considered by the GS.

Recall is calculated by the formula, $Recall = \frac{TP}{TP+FN}$, Precision is calculated by the formula $Precision = \frac{TP}{TP+FP}$, and F-measure, the harmonic mean of Precision and Recall, is calculated by the formula $F_1 = 2\ \frac{Precision*Recall}{Precision+Recall}$.

The Gold Standard consisted of grey literature reports containing instances of several entity types, including Archaeological Context (Feature), Artefact, Event, Material, Method, Monument, Place, Period and Person. The entities Event, Method, and Person were not in the scope of the NER pipeline and are not included in the evaluation.

The manually annotated GS was helpful for the purposes of an early evaluation task, revealing several issues with regards to vocabulary coverage and suggesting potential rule matching strategies for a range of different entities. The NER pipeline was evaluated against the GS delivering the performance figures as seen on Table 1. The overall score of Recall and Precision were encouraging, reaching 57% and 61% respectively and delivering an F-Measure score of 59%.

The lowest performing entity is Monument both in terms of Recall (36%) and Precision (45%), followed closely by Artefact which shares the same Recall score and slightly better (50%) Precision. The pipeline delivers slightly better scores for the Material and Place Entities with Recall scores between 50%-60% and Precision scores between 62%-65% respectively. The best performing entity is Archaeological Context which enjoys a Recall score of 87% and Precision 63%, followed by the Period entity which scores 72% Recall and 71% Precision. The contribution of vocabularies is critical to the performance of the pipeline with respect to the discussed entities. Clearly, Precision can be harmed by using too many terms from the available vocabulary which do not fail within the scope of the targeted entity. Conversely, Recall can be affected by using too few terms from the vocabulary.

The evaluation also revealed several issues related to vocabulary coverage and quality of the manually annotated corpus affecting precision and recall rates. Such cases concern missing, non-considered and out-of-scope annotations. Missing-annotations reflect the cases where matches are recorded as false positives instead of true positives due to the absence of such annotations from the GS. Manual annotation is a laborious and repetitive task that can impact the accurate annotation of every single instance of frequently occurring terms. Some instances may be overlooked, for example 'pit/ *kuil*' which is a frequently occurring term in archaeological grey-literature not having a polysemous behaviour. It is almost certain that every instance of this term missing from the GS is because it has been overlooked rather than purposely singled out.

Non–considered annotations concern a different set of cases than missing annotations, which are not overlooked during the manual annotation process but have not been addressed at all. Compared to the missing-annotation cases where some instances are included in the GS and some are overlooked, the non-considered cases fail to add in the GS all instances of a particular term. Most likely this is due to the manual annotation guidelines that do not scope such terms. For example, building/*gebouw*, field/*akker*, and road/*weg*, are not monuments in the strict sense but are useful terms for indexing worth to be addressed by the NER task. Similarly, out-of-scope annotations can be introduced in the GS following an open interpretation or misinterpretation of the guidelines by the annotators. The GS contained contemporary dates and non-Dutch place names which might be useful in other contexts but beyond the scope of our NER pipeline.

The evaluation results in combination with a closer examination of the Gold Standard suggested several alternative labels and synonyms for inclusion in the vocabulary. The size of the GS is not adequate to suggest a complete and comprehensive list. Whenever possible, we have used the GS to enrich the existing vocabulary with specialised terms and spelling variations. For example, we added alternative labels of spelling variations to period instances containing 'mid/*midden*' and 'late/*laat'*. Such periods moderators may appear in brackets e.g

'(Midden) Mesolithicum', and alterative labels were introduced for matching such period constructs. Other enrichments relate to specialised terms that appeared frequently in the GS and are synonyms of existing ontology classes. For example, the class labelled as 'seed/fruit/nut/kernel' (*zaad/vrucht/noot/pit*) which had already been split during the vocabulary enhancement stage to individual labels, has been enriched with the labels; grain/*graan*, micro-remains/*macroresten*), barley/*gerst* and wheat/*tawe*.

A new version of the Archaeological concepts pipeline was developed to incorporate the vocabulary modifications and a new set of rules was introduced for matching grid references of places and instances from the Landscape elements class of the *Objecttypen* facet thesaurus. The restriction that any match of a Place entity must commence with an upper-case letter has been lifted, to include matching for place names commencing with s', such as 's-Heerenberg, and 's-Graveland which is quite common in Dutch. Rules were improved for matching the date suffixes and combinations, enabling matching of date range such as between 1600 and 1900 (*tussen 1600 en 1900*). In addition, the pipeline incorporated new improved hand-crafted rules addressing polysemy of object/place entities for improving performance and strengthening the matching accuracy. For example, a rule aimed at matching instances of the artefact class, which previously included two conditions, was strengthened to include five separate conditions as seen below.

```
{!NotLookup, !Context, !Physical_Thing, Token.category == N,
Lookup.class == [rceArtefact n=ArcheoArtefactTypes]}
```

The rule matches all instances of the RCE Artefact Types class, excluding from matching; certain areas of the resource previously annotated as NotLookup, Archaeological Context, and Monuments (Physical Thing) whilst requiring each match to conform to Noun token-category.

The updated version was iteratively evaluated against the same GS (not previously unseen) delivering improved results (table 2).

The Recall of the NER pipeline is improved by 10% whereas Precision is also improved by 7%, reaching overall 68% Precision and 67% Recall. Most significantly, the performance of the pipeline has been considerably improved for the Artefact and Monument entities types, with Recall increasing from 36% to 53% and 64% respectively and Precision increasing to 63% and 65%. The Precision and Recall scores for Material have improved slightly – it should be remembered that the Material entity poses particular difficulties (for humans and machines) in the archaeological report context due to its ambiguity with Objects (see discussion below and in Vlachidis and Tudhope, 2016). The performance of the pipeline is comparable with the initial results of the AGNES system (Brandsen *et al.*, 2019) which employed a supervised ML approach for the recognition of Artefact, Time Period and Material entities in Dutch archaeological grey literature. Both systems reported comparable F-Measure scores for a range of entities and recognised the challenges imposed by the archaeological domain in NER. When systems have reached maturity, a full-scale comparative study across the rule-based, ML (supervised and unsupervised) methods of NER and potentially including a complementary use of both methods might be of interest. It is evident that the RCE Thesauri proved a valuable resource to drive the NER effort, providing a significant vocabulary breadth which benefited the Recall rates of the system. At the same time, the hand-crafted rules as

improved during the iterative process allowed for a maximum use of the available vocabulary whilst imposing conditions which protected the overall precision rates of the system.

6. Discussion

A major development of the NER rule-based and KOS-driven approach has been the generalisation of the previous rule-based techniques (Vlachidis and Tudhope, 2016) to Dutch archaeological grey literature. The work faced challenges in adapting to a different set of vocabularies available via the RCE Thesaurus and also to differences in language characteristics. The NER techniques were focused on the range of general archaeological and dendrochronological concepts and proved capable of extracting relevant entities of interest with relative success. The ABR and AAT Thesauri proved to be a valuable resource in support of the NER task, however, archaeological vocabularies do pose a challenge. Unlike highly specialised domains, which have vocabularies unique to that domain, archaeological terminology uses common words, for example “wall”, and “ditch” in a very particular context (see discussion in Vlachidis and Tudhope, 2012). In addition, thesaurus resources are typically developed with Information Science principles in mind and might require adaptation to NLP tasks. The Gold Standard (GS) evaluation revealed performance drawbacks influenced by structural, labelling and coverage issues of the vocabulary. The results of the evaluation phase led to resolving the overloaded vocabulary entries into individual term components. Labelling adjustment and enrichment techniques are necessary for making the vocabulary resources applicable to the NER focus as discussed in Section 3.

The performance of the Dendrochronological concepts pipeline has not been quantitatively evaluated due to the lack of a gold standard at the time of development. The results have been qualitatively inspected confirming useful output which has been disseminated to project partners within the ARIADNE Infrastructure. The current version of the pipeline is limited to assertion of generic weightings on sentences containing two or three different entity types, Future work should apply more contextualised information extraction, building on techniques from previous work on English relation extraction (Vlachidis and Tudhope, 2016). Such contextualised extractions are based on grammatical patterns for identifying relationships between objects and dates or material. The semantic annotations produced by the Dendrochronology pipeline are assigned URI references from the AAT thesaurus, which has been used as the backbone vocabulary of ARIADNE for linking multilingual concepts. Such output has been used to demonstrate the principle of data integration across multilingual archaeological grey literature and excavation databases using concepts aligned to AAT (Binding *et al.*, 2019).

Vocabulary coverage can impact recall of rule-based systems where absence even of a single concept from the vocabulary that frequently occurs in text can significantly harm recall rates. Such recall performance issues can be propagated to a ML setting should the same vocabulary be employed or should the rule-based output be used for training purposes. The rules of the NER pipeline were designed to reveal entities of interest and among other uses to aid discovery and metadata generation. Therefore, quantifiers were included in the annotation spans to enrich the semantics of an extracted entity. This was particularly evident with Time Period entities where a qualifier significantly adds to the meaning, for example “roughly around 200BC” is semantically a richer annotation in the archaeological sense than simply

"200BC". Such semantically rich spans that include qualifiers can be useful for metadata and discovery purposes.

The Archaeological concepts pipeline has been challenged by a series of domain and language-oriented aspects. A word-sense ambiguity specifically relevant to the archaeology domain concerns material and object senses, and appears in English and Dutch, and most likely in other languages (Vlachidis 2012 ; Bransden *et al.*, 2020). Under the expert eye of archaeologists, a very small piece of material e.g., pottery (*aardewerk* in Dutch) can be treated as a find. Hence, the boundary between material and physical object blurs, which can be particularly challenging from a NER perspective where instances are expected to be classified under a single category. To the human investigator the distinction between the two senses may be contextually evident but from a computationally perspective this fine distinction is hard to detect.

A particular language related challenge of NER task relates to the way nouns are synthesised in Dutch to produce compound noun forms. Such compound nouns in the context of archaeological reports typically join period with object and objects with material, for example, 'pottery fragment / *aardewerkfragment*'. Employing partial matching Lookup instead of whole word matching carries the potential of addressing annotation of compound nouns. Partial matching is possible in GATE but should be planned and executed carefully due to the significant amount of noise that can be generated and complications in entity type assignment. A choice should be made to decide whether such cases are annotated as single entity spanning across the compound nouns or two separate annotations are assigned to each part of the entity. Further work is needed for addressing the issue of compound words comprehensively both on the practical pattern matching and on the annotation assignment levels.

A future improvement of the pipeline should also be able to address negated entities that provide facts of no evidence, i.e. a comment in the report that no evidence has been found for a potential finding and thus should not be annotated. Both NER pipelines (Archaeological concepts and Dendrochronological concepts) ignore negated finds (e.g., no finds / *geen vondsten*) and deliver false-positive annotations for non-affirmative instances in text. Negation detection of archaeological entities in English has been addressed by a previous study that employed domain glossary and hand-crafted rules to detect phrases of negated facts (Vlachidis and Tudhope, 2015). An equivalent module could be incorporated in a future version to enable detection of negated instances in Dutch archaeological grey literature.

Restricting entity recognition to nouns supports precision but at the same time limits recall on adjectival forms, many of which have been identified as relevant by the gold standard, such as bronzen (bronze), stenen (stone) etc. Future version should revisit this restriction and plan for rules which could approach such instances as individual material entities or as moderators of object or monument entities. In addition, identifying passages of particular relevance for information extraction can improve the focus and precision of results, given the length of many archaeological reports. In the case of Dendrochronology NER the section detection approach proved valuable. Future versions could expand the approach to detect sections that can be prioritised, omitted or processed with specialised NLP components. The variety of report structure and styles can be a hindrance to accurate detection and practical

approaches could be attempted for focusing on sections that merit unique characteristics (e.g. abstracts).

7. Conclusions

The details of a rule-based, KOS driven NER method that led to the development of two NLP pipelines, targeted respectively at the recognition of general archaeological and of specifically dendrochronological concepts, in Dutch grey literature have been discussed. The work has been motivated by the need to enable access and facilitate discovery of archaeological knowledge currently untapped and obscure within grey-literature. The pipelines were developed in the GATE framework and are available as standalone GATE applications from the ARIADNE services portal and as cloud web services from the GATE Cloud. The employment of domain vocabulary has been critical for driving the NER task and for enabling assignment of terminological references to the extracted entities with respect to domain thesauri. A range of vocabulary adjustment and enrichment processes have been followed for adapting labels of thesaurus terms to the NER task following natural language descriptions. The result of archaeological concept NER is evaluated against a Gold Standard, human-annotated corpus. The NER task is challenged by a series of domain and language-oriented aspects, concerning vocabulary coverage, complex forms of compound nouns and negations. The results suggest the suitability of rule-based KOS driven approaches for attaining the low-hanging fruits of NER using a combination of quality vocabulary with a small set of rules. Their generalisability can be further confirmed using an unseen evaluation set which at the point of development was not available. The NER techniques were focused on general archaeological and dendrochronological entities and the method proved capable of extracting entities of interest with relative success. Future improvements include extraction of compound noun forms which appear in Dutch regularly that join entities of interest together to create complex noun form of period with objects, material, archaeological contexts etc. Being able to detect negated facts and entities in text is also another future direction that can improve the NER method which can further benefit from section detection approaches for fine tuning and focusing on particular document areas.

References

Amrani, A., Abajian, V., Kodratoff, Y. and Matte-Tailliez, O. (2008), 'A chain of text-mining to extract information in archaeology', in *2008 3rd International Conference on Information and Communication Technologies: From Theory to Applications*, IEEE, Damascus, Syria, pp. 1-5, doi: 10.1109/ICTTA.2008.4529905

Angjeli, A., Mac Ewan, A. and Boulet, V. (2014), 'ISNI and VIAF – Transforming Ways of Trustfully Consolidating Identities', Paper Presented at the *IFLA WLIC 2014 - Lyon - Libraries, Citizens, Societies: Confluence for Knowledge in Session 86 - Cataloguing with Bibliography, Classification & Indexing and UNIMARC Strategic Program*me. 16-22 August 2014, Lyon, France. Available from http://library.ifla.org/id/eprint/985/1/086-angjeli-en.pdf (Accessed 10 September 2021).

Art and Architecture Thesaurus. [online] Available at: http://www.getty.edu/research/tools/vocabularies/aat/about.html (Accessed 10 September 2021)

Binding, C., Tudhope, D. and Vlachidis, A. (2019) 'A study of semantic integration across archaeological data and reports in different languages', *Journal of Information Science*, Vol.45 No.3, pp.364-386.

Bontcheva, K., Tablan, V., Maynard, D. and Cunningham, H. (2004) 'Evolving GATE to meet new challenges in language engineering', *Natural Language Engineering*, Vol.10 No.3-4, pp.349-373.

Brandsen, A., Lambers, K., Verberne, S. and Wansleeben, M. (2019) 'User Requirement Solicitation for an Information Retrieval System Applied to Dutch Grey Literature in the Archaeology Domain', *Journal of Computer Applications in Archaeology*, Vol.2 No.1, pp.21-30

Brandsen, A., Verberne, S., Wansleeben, M. and Lambers, K. (2020) 'Creating a Dataset for Named Entity Recognition in the Archaeology Domain', in *Proceedings of the 12th Language Resources and Evaluation Conference*, The European Language Resources Association, Marseille, France, pp. 4573-4577.

Bunn, A.G. (2008) 'A dendrochronology program library in R (dplR)', *Dendrochronologia*, Vol. *26 No.*2, pp.115-124.

Byrne, K.F., Klein, E. (2010) 'Automatic Extraction of Archaeological Events from Text', in B. Frischer, J.W. Crawford and D. Koller (Eds.) *Making History Interactive. Proceedings of the 37th Computer Application in Archaeology Conference*, BAR Publishing, Williamsburg, Virginia US, pp.48–56

Cunningham, H., Tablan, V., Roberts, A. and Bontcheva, K. (2013). 'Getting more out of biomedical documents with GATE's full lifecycle open source text analytics', *PLoS Computational Biology*, Vol.9 No.2: e1002854.

Erfgoedthesaurus. [online] Available at: https://thesaurus.cultureelerfgoed.nl/ (Accessed 10 September 2021)

Evans, T.N.L. (2015) 'A Reassessment of Archaeological Grey Literature: semantics and paradoxes', *Internet Archaeology,* Vol. 40 [online] https://doi.org/10.11141/ia.40.6 (Accessed 10 September 2021)

Farace, D.J. (1997) 'Grey Literature and Publishing', *Publishing Research Quarterly,* Vol. 13 No.2, pp. 3-4.

Fiorucci, M., Khoroshiltseva, M., Pontil, M., Traviglia, A., Del Bue, A. and James, S. (2020) 'Machine learning for cultural heritage: A survey', *Pattern Recognition Letters,* Vol. 133, pp.102-108.

Garoufallou, E. and Ovalle-Perandones, M.A. eds., (2021) *Metadata and Semantic Research: 14th International Conference, MTSR 2020, Madrid, Spain, December 2–4, 2020*, Revised Selected Papers. Proceedings. Communications in Computer and Information Science (CCIS), Vol. 1355, pp. XXIV, 412. Springer Nature, Cham, Switzerland.
DOI: https://doi.org/10.1007/978-3-030-71903-6

Gelling, M. (2011) 'Place-Names and Archaeology', in Hinton, D. A., Crawford, S. and Hamerow, H. (Eds), *The Oxford Handbook of Anglo-Saxon Archaeology* [online]. Oxford University Press doi:10.1093/oxfordhb/9780199212149.013.0050 (Accessed 10 September 2021)

Grishman, R. and Sundheim, B.M. (1996) Message understanding conference-6: A brief history, in *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, Center for Sprogteknologi, Copenhagen, Denmark, pp. 466–471.

Jansma, E., van Lanen, R., Brewer, P. and Kramer R. (2012) 'The DCCD: A digital data infrastructure for tree-ring research', *Dendrochronologia*, Vol. 30, No. 4, pp. 249-251.

Jeffrey, S., Richards, J., Ciravegna, F., Waller, S., Chapman, S. and Zhang, Z. (2009) 'The Archaeotools project: faceted classification and natural language processing in an archaeological context', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, Vol. 367 No.1897, pp.2507-2519.

Kraaij, W., and Pohlmann, R. (1994) 'Porter's stemming algorithm for Dutch', in Noordman LGM and de Vroomen WAM (Eds), *Informatiewetenschap 1994: Wetenschappelijke bijdragen aan de derde STINFON Confere*ntie, Stichting Informatiewetenschap Nederland Tilburg, Netherlands, pp. 167-180.

Kuniholm, P.I. (2002) 'Archaeological dendrochronology', *Dendrochronologia*, Vol. 20 No.1-2, pp.63-68.

Meghini, C., Scopigno, R., Richards, J., Wright, H., Geser, G., Cuy, S., Fihn, J., Fanini, B., Hollander, H., Niccolucci, F. and Felicetti, A. (2017) 'ARIADNE: A research infrastructure for archaeology', *Journal on Computing and Cultural Heritage* (JOCCH), Vol. 10 No. 3, pp.1-27.

Nadeau, D. and Sekine, S. (2007) 'A survey of named entity recognition and classification', *Lingvisticae Investigationes*, Vol. 30 No.1, pp.3-26.

Paijmans, H., Wubben, S. (2007) Preparing archaeological reports for intelligent retrieval, in A. Posluschny, K. Lambers and I. Herzog (Eds.) *Layers of Perception. Proceedings of the 35th International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA)*, Habelt Verlag , Berlin, Germany, pp. 212–217

Piskorski, J., Wieloch, K. and Sydow, M. (2009) 'On knowledge-poor methods for person name matching and lemmatization for highly inflectional languages', *Information Retrieval*, Vol. 12 No. 3, pp.275-299.

Piskorski, J. and Yangarber, R. (2013) 'Information extraction: Past, present and future' in *Multi-source, multilingual information extraction and summarization*, Springer, Berlin, Germany, pp. 23-49.

Richards, J. and Hardman, C. (2008) 'Stepping back from the trench edge: an archaeological perspective on the development of standards for recording and publication', in Greengrass, M. and Hughes, L. (Eds.) *The Virtual Representation of the Past*, Routledge, Farnham England, pp. 101-112

Rijksdienst vvor het Cultureel Erfgoed. (2019) *De Erfgoedmonitor*. Available at: https://erfgoedmonitor.nl/indicatoren/archeologisch-onderzoek-aantal-onderzoeksmeldingen (Accessed 10 September 2021)

Selhofer, H. and Geser, G. (2014) *D2.1: First Report on Users' Needs. Technical report.* [online] Project Report ARIADNE Infrastructure. http://legacy.ariadne-infrastructure.eu/wp-content/uploads/2019/07/ARIADNE_D2- 1_First_report_on_users_needs.pdf (Accessed 10 September 2021)

Toledo, J.I., Carbonell, M., Fornés, A. and Lladós, J. (2019) 'Information extraction from historical handwritten document images with a context-aware neural model', *Pattern Recognition*, Vol. 86, pp.27-36.

Tudhope, D., May, K., Binding, C. and Vlachidis, A. (2011) 'Connecting archaeological data and grey literature via semantic cross search', *Internet Archaeology*, Vol. 30 [online] https://doi.org/10.11141/ia.30.5 (Accessed 10 September 2021)

Van Hooland, S., De Wilde, M., Verborgh, R., Steiner, T. and Van de Walle, R. (2015) 'Exploring entity recognition and disambiguation for cultural heritage collections', *Digital Scholarship in the Humanities*, Vol. 30 No.2, pp.262-279.

Vlachidis A., Tudhope D., Wansleeben M. (2021) 'Knowledge-Based Named Entity Recognition of Archaeological Concepts in Dutch', in: Garoufallou E., Ovalle-Perandones MA. (Eds) *Metadata and Semantic Research. MTSR 2020. Communications in Computer and Information Science*, vol 1355, pp. 53-64, Springer, Cham. DOI: https://doi.org/10.1007/978-3-030-71903-6_6

Vlachidis, A. and Tudhope, D. (2016) 'A knowledge-based approach to Information Extraction for semantic interoperability in the archaeology domain',*Journal of the association for information science and technology*, Vol.67 No.5, pp.1138-1152.

Vlachidis, A., and Tudhope, D. (2015) 'Negation detection and word sense disambiguation in digital archaeology reports for the purposes of semantic annotation', *Program: Electronic Library and Information Systems,* Vol.49 No.2, pp. 118-134

Vlachidis, A., and Tudhope, D. (2012) 'A pilot investigation of information extraction in the semantic annotation of archaeological reports', *International Journal of Metadata, Semantics and Ontologies*, Vol. 7 No.3, pp.222-235.

Vlachidis, A. (2012) *Semantic indexing via knowledge organization systems: Applying the CIDOC-CRM to archaeological grey literature.* PhD thesis, University of Glamorgan, United Kingdom

Figure 1 Example results of the Archaeological Concepts NER pipeline: Different entity types are annotated in colours; yellow monuments (e.g. nederzettingen), blue Period (e.g. Middeleeuwen), orange Artifact/Find (e.g. houtskool), purple Material (hout), brown Archaeological Context (kuilen) and green Place (Oost-Nederland)

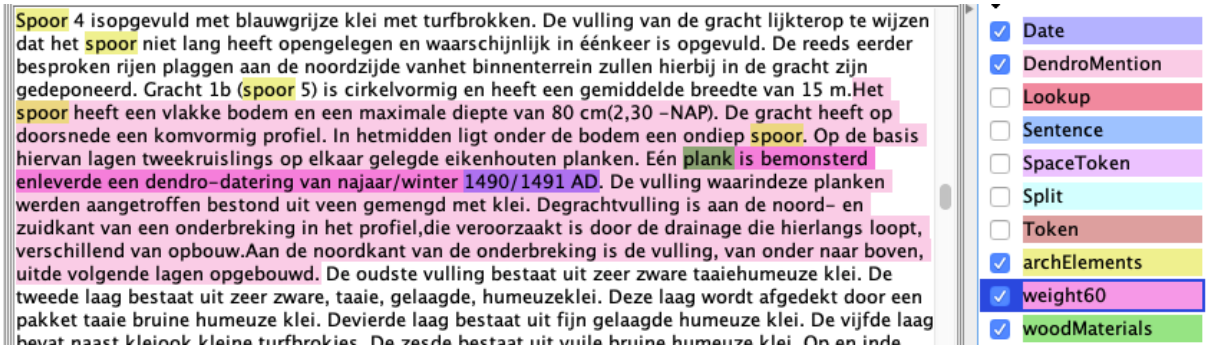

Figure 2 Example results of the Dendrochronological Concepts NER pipeline: Different entity types are annotated in colours. Both individual entities and longer text passages are shown ;yellow architectural elements (e.g. spoor), blue Date (e.g. 1490/1491 AD)), purple weighted phrase and pink dendrochronology discussion section.

Table 1. System Performance of the early NER system for a range of entities

Entity	Recall	Precision	F-Measure
Arch.Context	0.87	0.63	0.73
Artefact	0.36	0.50	0.42

Material	0.50	0.62	0.55
Monument	0.36	0.45	0.40
Place	0.60	0.65	0.63
Time Period	0.72	0.70	0.71
All(Total)	**0.57**	**0.61**	**0.59**

Table 2. System Performance of the updated NER system for a range of entities

Entity	**Recall**	**Precision**	**F-Measure**
Arch.Context	0.90	0.63	0.74
Artefact	0.53	0.63	0.56
Material	0.53	0.63	0.57
Monument	0.64	0.65	0.65
Place	0.61	0.73	0.67
Time Period	0.80	0.77	0.79
All(Total)	**0.67**	**0.68**	**0.67**