2016/17 Data Analytics Project Lot 1: Measurement and Evaluation of a Gold Dataset for Text Processing

Dataset Requirement Framework Specification

1. Introduction

This technical memo has been produced as a deliverable associated with Lot 1 (Measurement and Evaluation of a Gold Dataset for Text Processing) of the RCloud Statement of Requirement for Dstl's 2016/17 Data Analytics project. The memo has been written by Aleph Insights Limited, with inputs from Tenode Limited.

Lot 1 calls for the development of a 'Gold Standard' dataset which can subsequently be used for training and validating machine learning approaches to natural language processing and text analytics in a defence and security context. This document provides a detailed description of the dataset requirement framework (DSRF), which forms the basis for the content and the structure of this gold standard dataset.

The DSRF comprises two main elements:

1. The set of criteria used to select data sources and documents for inclusion within the dataset;
2. The schema used to tag both entities and relationships in the text of the documents included in the dataset.

This document provides both a theoretical justification of the approaches taken within the DSRF, and a detailed explanation of their execution with regards to the compilation of the gold standard dataset. It includes the following sections:

- **The Requirement** - an exposition of the need for the gold standard dataset and an articulation of the project aim.
- **Theoretical Basis** - a justification for the DSRF founded on existing research and theory.
- **DSRF Methodology** - a description of the method employed to select sources and documents for the gold standard dataset, the structural schema used to tag entities and the relationships between them, and the approach that will be employed to measure confidence in the quality of the tagging.

2. The Requirement

Intelligence analysts operating in the defence and security domain are required to digest, comprehend and make inferential links across ever expanding datasets. The amount of unstructured information contained within the vast array of intelligence reports, open source material, social media data and other sources that are available on any given topic, far exceeds the human resource allocated by government to the analysis of that topic.

The existence of research projects such as the Asymmetric Threat Response and Analysis Program (ATRAP)[1] are an acknowledgement that there is a pressing need to apply sophisticated automated solutions to the analysis of large relevant datasets, in order to help intelligence analysts to fulfil their role in maintaining Defence's information advantage.

At the most basic level, these solutions must be capable of reliably and consistently extracting entities (e.g. people, organisations and locations) from a wide range of document types and capturing the relationships between the extracted entities (e.g. whether they are co-located or have communicated with one another). This will necessitate the further development of machine learning algorithms which extract structured information from unstructured sources (including those within the Baleen text analysis framework developed by Dstl[2]).

In order to develop, train, measure and validate these machine learning approaches - such that they can be demonstrated to fulfil this function without human input, to a sufficient standard and across a broad selection of defence tasks - there is the requirement to develop a gold standard[3] dataset containing documents which are accurately and comprehensively tagged.

A number of such datasets exist, notably the Brown Corpus, the Penn Treebank sample and the 2003 Conference on Computational Natural Language Learning (CoNLL) dataset[4]. Most of these datasets are tagged according to either generic grammatical and syntactical schema, or esoteric schema relating to particular fields (e.g. medicine). As such, they lack the specific tagging structure necessary to encode the defence-relevant information required to support an intelligence analyst in their job.

Furthermore, the content of these existing datasets is not representative of the datasets encountered by intelligence analysts in the real world (i.e. they are lacking in content validity as described by Marrero et al.[5] (2012)). Whilst the Message Understanding Conferences (MUC), supported by the Defense Advanced Research Project Agency (DARPA), created more relevant annotated datasets[6], these were tagged with their own bespoke schema developed by the conference participants to meet the conference organisers' requirements.

---

[1] Chan, E., Ginsburg, J., Ten Eyck, B., Rozenblit, J. and Dameron, M., 2010, May. Text analysis and entity extraction in asymmetric threat response and prediction. In Intelligence and Security Informatics (ISI), 2010 IEEE International Conference on (pp. 202-207). IEEE.

[2] https://github.com/dstl/baleen/wiki/An-Introduction-to-Baleen

[3] There is no universally accepted definition of a gold standard dataset for NLP, but it is generally considered to be a dataset which has been extensively tagged by a range of human annotators whose outputs are cross-validated against one another, and against 'objective' automated tagging; inter-annotator agreement is subsequently calculated to ensure quality. Wissler, L., Almashraee, M., Díaz, D.M. and Paschke, A., 2014. The Gold Standard in Corpus Annotation. In IEEE GSC.

[4] An extensive list of tagged datasets has been produced by the Natural Language Toolkit (NLTK) and can be found here: http://www.nltk.org/nltk_data/

[5] Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J., and GómezBerbás, J. M. Named entity recognition: Fallacies, challenges and opportunities. Journal of Computer Standards and Interfaces, 32(5):482–489, 2012.

[6] Grishman, R. and Sundheim, B., 1996, August. Message Understanding Conference-6: A Brief History. In COLING (Vol. 96, pp. 466-471).

The aim of this project, therefore, is to develop a gold standard dataset which includes a collection of documents that are germane to the defence and security domain, and which are tagged using task-specific entity and relationship schema. The dataset will be designed to reflect the types of data that are currently processed by human analysts studying defence-specific questions, and which it is hoped will be analysed effectively by automated extraction tools in the future.

3. Theoretical Basis

Compiling and annotating a dataset to meet the requirement described above will require an appreciation of the challenges that are faced by ML algorithms as they attempt to tag natural language text. This section considers the primary linguistic challenges faced when trying to infer and encode meaning within natural language, in order to ensure that the dataset is optimally structured to meet the requirement.

3.1. Named Entity Recognition and Classification

The first difficulty encountered in tagging text for its meaning is presented by the concept of named entity recognition (NER). NER is the process of identifying named entities within text and allocating them to predefined (or sometimes emergent) categories. Its aim is to identify and classify the proper names[7] (entities) within a given text. This is an essential step in the process of deconstructing textual meaning, and is a necessary precursor for considering the interactions and relationships between entities[8]. The classification system or schema used to identify and categorise entities within the text will ultimately determine what information can be extracted from any text.

Generally the entity categories within tagging schema include persons, places and organisations, but they can also include other kinds of specific nouns (e.g. vehicles) or other types of properties (e.g. quantities or times). The Baleen entity schema, which includes all of these, was developed prior to this project and its use within this project is detailed in section 4.5.

3.2. Relationship Extraction

Relationship extraction (RE) refers to the task of understanding how different extracted entities relate to one another. There are a number of different conceptual relationships which might be considered in this context, including:

- Hierarchical relationships - using a taxonomic structure to define hierarchies for sets and subsets of terms[9];

[7] In this context, a proper name should be considered to be the words or string of words which define a given referent.

[8] Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J. and Gómez-Berbís, J.M., 2013. Named entity recognition: fallacies, challenges and opportunities. Computer Standards & Interfaces, 35(5), pp.482-489.

[9] Wang, T., Li, Y., Bontcheva, K., Cunningham, H. and Wang, J., 2006, June. Automatic extraction of hierarchical relations from text. In European Semantic Web Conference (pp. 215-229). Springer Berlin Heidelberg. and Neelakantan, A., Passos, A. and McCallum, A., A Hierarchical Model for Universal Schema Relation Extraction.

- Grammatical relationships - where syntactical interaction between the words is considered[10], including part-of-speech tagging (as conducted in the Brown Corpus);
- Information relationships - examining the structures used to store information as used in relational databases[11];
- Causal relationships - where predictive interactions between entities are studied, commonly involving the examination of medical texts for links between factors, treatments and disease[12].

Numerous attempts have been made to develop universal relationship schema[13]; however, RE is often driven by specific intelligence requirements and thus bespoke domain and task-specific schema are used to structure the analysed text. Intelligence analysts are primarily concerned with individuals, organisations and objects of military relevance (such as weapons and vehicles), and their temporal and geographic situation[14]. They are also interested in information regarding the affiliations of these entities, their sentiment towards one another and their level of cooperation. The relationship schema developed for this project, which is described in detail in Section 4.6 of this document, has been designed to capture this task-specific information.

The structure of this relationship schema deliberately contains as few categories and as few classifications per category as is possible while still maintaining usefulness, in order to limit the relationship instances in the dataset to a manageable number. Given all of the relationship types are n-ary in nature (i.e. can apply to multiple combinations of entities simultaneously), the potential for the exponential growth of relationship instances is high. Additionally, the broader the range of relationships included in the schema, the greater the probability that multiple relationships will apply between the same pair or set of entities, further compounding the potential number of relationship instances. An imperative exists, therefore, to restrict any superfluous relationship types.

It should be noted that relationships can take a number of different forms. The three main forms of relationship applicable to the type of relationship schema employed in this project are: directional, symmetric and transitive.

Directional relationships, such as the relation of 'belonging to', imply an inherent direction. The sentence - "The car belongs to the man" - tells us that the man has some legal or historical claim of ownership over the car, but the inverse is not true: the man does not belong to the car. Directional relationships can also be reciprocal and apply in both directions, as with the relation 'to like'. For

[10] Jurafsky, D. and Martin, J.H., 2000. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.

[11] Andersson, M., 1994, December. Extracting an entity relationship schema from a relational database through reverse engineering. In International Conference on Conceptual Modeling (pp. 403-419). Springer Berlin Heidelberg. and Al-Masree, H.K., 2015. Extracting Entity Relationship Diagram (ERD) from relational database schema. International Journal of Database Theory and Application, 8(3), pp.15-26.

[12] Ramakrishnan, C., Kochut, K.J. and Sheth, A.P., 2006, November. A framework for schema-driven relationship discovery from unstructured text. In International Semantic Web Conference (pp. 583-596). Springer Berlin Heidelberg.

[13] Hachey, B., 2009. Towards generic relation extraction.

[14] Chan, E., Ginsburg, J., Ten Eyck, B., Rozenblit, J. and Dameron, M., 2010, May. Text analysis and entity extraction in asymmetric threat response and prediction. In Intelligence and Security Informatics (ISI), 2010 IEEE International Conference on (pp. 202-207). IEEE.

example: person A can like person B without reciprocation; person B can like person A without reciprocation; or they can both like each other.

Symmetric relationships are those which by necessity apply to both entities at the same time. The relation 'communicated with' is an example of a symmetric relationship, where it is necessary for both parties to be involved in an interchange for a communication to have happened[15]. 'Co-located with' is also an example of a symmetric relationship. The sentence, "The man sat in the car" - tells us that the man AND the car are in the same place, and neither can be elsewhere while the other is co-located with it. However, 'co-located with' can also be transitive, where a relationship between two entities is implied by their own separate relationships with a third entity. In the sentence, "The man sat in the car, which was parked in the carpark" - the man and the car are explicitly co-located, and the car and the carpark are explicitly co-located, ergo the man and the carpark must be co-located. Understanding the form of a relationship is central to capturing the meaning that arises from the interaction between entities within a text.

In essence, the combination of the entity and relationship schema leads to a simplified representation of the 'meaning' of the text and allows this distilled and structured information to be extracted[16]. The tagging system thus defines the information content which can be extracted from text. The design of the tagging schema is therefore fundamental to determining what can be 'known' about the dataset, and what future ML approaches can learn from the dataset or be tested against.

The schemas used in this project are generic, in that they are designed to cover the most common types of intelligence question; however, they are not universal. Therefore, ML systems trained on the dataset, which is based on these schema, will also be non-universal.

3.3. Challenges in Information Extraction

Having designed an appropriate schema, there are still a number of conceptual challenges to be overcome in order to effectively tag a dataset. Aside from the more abstract task of determining what is meant at the foundational level by any given referent (the word, phrase or set of text associated with a specific entity), as discussed notably by Frege[17] and Kripke[18], there are a number of practical challenges resulting from the attempt to tag entities and relationships.

Many of these challenges stem from difficulties in applying knowledge external to the given text to provide context and additional understanding of meaning, and therefore facilitate the identification and labelling of an entity or a relationship. Humans regularly apply knowledge obtained from outside the text under consideration to add information and infer meaning[19]. The transference of knowledge

[15] In this context a transmission of a message by entity A which is not received by entity B is considered to constitute a failed communication.

[16] Collobert, Ronan, et al. "Natural language processing (almost) from scratch." Journal of Machine Learning Research 12.Aug (2011): 2493-2537.

[17] Frege, G., 1948. Sense and reference. The philosophical review, 57(3), pp.209-230.

[18] Kripke, S.A., 1972. Naming and necessity. In Semantics of natural language (pp. 253-355). Springer Netherlands.

[19] Rouet, J.F. and Britt, M.A., 2011. Relevance processes in multiple document comprehension. Text relevance and learning from text, pp.19-52.

represents a considerably more difficult problem for machine learning approaches to natural language processing (NLP)[20], particularly when moving across subject domains[21]. The utility of external knowledge to NLP is recognised in part by the widespread use of gazetteers for improving extraction of named entities[22]. However, gazetteers purely represent additional vocabulary, and do not assist with many of the cognitive and inferential tasks performed by human readers drawing on external knowledge. This also does not help much with the extraction of relationships from text, where a given relationship class can be expressed in a multitude of different ways.

Inference skills are widely deployed by human readers to extract meaning from text, which is not explicitly present in the text itself. This opaque process, which is presumed to draw on complex cognition, prior knowledge and a host of learnt semantic and syntactic clues[23], allows the human reader to identify and classify entities and to understand the relationships between them.

Some of this capability can be replicated by the use of rules-based solutions (e.g. proper nouns beginning with capital letters), the utilisation of regular expression (regex) catalogues such as Stanford's RegexNER[24], the application of thematic ontologies[25] and the deployment of lexical databases, most notably WordNet[26]. However, automated entity extraction still underperforms human annotation in most cases[27], and automated RE trails human performance to an even greater degree[28].

The accuracy of automated information extraction deteriorates even further when parsing is scaled up from the sentence level, to the intra-document level or even across multiple documents[29]. One of the greatest problems posed by conducting entity and relationship extraction beyond the structure

[20] Wang, B., Guo, S., Liu, K., He, S. and Zhao, J., 2016. Employing External Rich Knowledge for Machine Comprehension. In Proceedings of IJCAI.

[21] Ciaramita, M. and Altun, Y., 2005. Named-entity recognition in novel domains with external lexical knowledge. In Proceedings of the NIPS Workshop on Advances in Structured Learning for Text and Speech Processing.

[22] Kazama, J.I. and Torisawa, K., 2007, June. Exploiting Wikipedia as external knowledge for named entity recognition. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL) (pp. 698-707).

[23] Kispal, A., 2008. Effective teaching of inference skills for reading: Literature review.

[24] http://nlp.stanford.edu/software/regexner.html

[25] Pisarev, I.A. and Kotova, E.E., 2016, February. Construction of thematic ontologies using the method of automated thesauri development. In 2016 IEEE NW Russia Young Researchers in Electrical and Electronic Engineering Conference (EIConRusNW) (pp. 305-307). IEEE.

[26] https://wordnet.princeton.edu/

[27] Nadeau, D. and Sekine, S., 2007. A survey of named entity recognition and classification. Lingvisticae Investigationes, 30(1), pp.3-26.; Baluja, S., Mittal, V.O. and Sukthankar, R., 2000. Applying Machine Learning for High-Performance Named-Entity Extraction. Computational Intelligence, 16(4), pp.586-595. and Marrero, M., Sánchez-Cuadrado, S., Lara, J.M. and Andreadakis, G., 2009. Evaluation of named entity extraction systems. Advances in Computational Linguistics, Research in Computing Science, 41, pp.47-58.

[28] Nguyen, T.H. and Grishman, R., 2015, June. Relation extraction: Perspective from convolutional neural networks. In Proceedings of NAACL-HLT (pp. 39-48).

[29] Masterson, D. and Kushmerick, N., 2003, September. Information extraction from multi-document threads. In ECML-2003: Workshop on Adaptive Text Extraction and Mining (pp. 34-41).

of a single sentence is the extension of coreference[30], namely trying to track multiple references to the same entity across a text[31] or multiple texts[32] (e.g. multiple references to 'Bashar Al Assad' across a range of texts, where Arabic transliterations for his name may differ, or he may be referred to as the 'President of Syria', or simply 'he'). The longer the span of text being dealt with the more taxing the issue of coreference becomes, as the number of entities 'at play' increases and creates confusion and disambiguation challenges[33]. Machine learning approaches particularly struggle with bridging anaphora[34], where coreference to an entity is often carried across sentences by fairly abstract and indirect linguistic mechanisms[35].

Therefore, within this current project, coreference of extracted entities and the tagging of the relationships between them will only be conducted at the local level (i.e. within single sentences). Global coreference and relationship attribution (i.e. for whole documents or across multiple documents) is considered a higher order task beyond the scope of the development of this dataset.

3.4. Corpus Data Structure

Even given the decision to limit tagging to sentence level structures, ambiguity within the annotation of the dataset is still highly likely. Natural language is inherently ambiguous. Allusion, metaphor, sarcasm, irony, double entendre, euphemism and many other rhetorical and literary devices deliberately create ambiguity - intended ambiguity; however, ambiguity can also arise when a reader interprets text to derive a meaning different to the one intended by the author - unintended ambiguity.

Semanticians have argued over whether true meaning stems from the design of the author, the interpretation of the audience or even some set of prescriptive and objective rules about language and definition[36]. Regardless of this debate, in practical terms the use of natural language gives rise to a level of uncertainty. For the purposes of this project, this uncertainty shall be defined as ambiguity.

[30] Zelenko, D., Aone, C. and Tibbetts, J., 2004. Coreference resolution for information extraction. In Proceedings of the ACL Workshop on Reference Resolution and its Applications (pp. 9-16).

[31] Known as 'document coreference'.

[32] Known as 'corpus coreference'.

[33] Huang, J., Taylor, S.M., Smith, J.L., Fotiadis, K.A. and Giles, C.L., 2009, June. Solving the Who's Mark Johnson puzzle: information extraction based cross document coreference. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Student Research Workshop and Doctoral Consortium (pp. 7-12). Association for Computational Linguistics.

[34] Hou, Y., Markert, K. and Strube, M., 2014, October. A Rule-Based System for Unrestricted Bridging Resolution: Recognizing Bridging Anaphora and Finding Links to Antecedents. In EMNLP (pp. 2082-2093).

[35] See this article for a list of types of bridging anaphora: Mineur, A.M., 2006. The resolution of bridging anaphora in OT.

[36] Examples of these complex arguments include: Tarski, A., 1944. The semantic conception of truth: and the foundations of semantics. Philosophy and phenomenological research, 4(3), pp.341-376. ; Kripke, S.A., 1972. Naming and necessity. In Semantics of natural language (pp. 253-355). Springer Netherlands. ; Frege, G., 1948. Sense and reference. The philosophical review, 57(3), pp.209-230.

The existence of ambiguity in natural language poses a problem for the construction of a gold standard dataset, which is assumed to contain the 'correct' annotation for entities and relationships. The interpretation of the relationship between two or more entities is often a highly subjective analytical process, dependent on the assessor's internal definition of the relationship term. The same is true to a lesser extent for the task of entity classification, where linguistic phenomena such as metonymy[37] create uncertainty about the category in which an entity should be placed.

Annotation is, therefore, often dependent on an individual's interpretation of the extraction schema, and the question of 'correct' annotation is subject to uncertainty. The greater the extent of this uncertainty, the greater the ambiguity of the text in question can be said to be.

The concept of textual entailment provides a way to accept and capture this ambiguity within a tagged dataset. Textual entailment was developed as an approach to aid natural language inference[38]. It involves generating a hypothesis (h) about the meaning of a given piece of text (t) and then assessing the probability of this hypothesis being true.

In the case of entity recognition and classification, it could be used to pose a question about whether an entity has been recognised and classified correctly, as in the following example:

> t: **President Obama** met with his Chinese counterpart.
>
> ⇒ h: President Obama is a PERSON

This approach could also be applied to relationship extraction, where a hypothesis about the relationship between different entities is postulated[39]. For example:

> t: President Obama met with his Chinese counterpart.
>
> ⇒ h: President Obama was co-located with his Chinese counterpart.

Textual entailment provides a model for measuring the ambiguity of the application of the tagging schema by assessing the probability that the hypothesis is true. The higher the assessed probability, the lower the ambiguity of the meaning of that instance should be considered to be.

In most cases this assessment of probability is determined by the level of inter-annotator agreement. The greater the agreement and the greater the number of those agreeing, the greater the probability that the inference represented by the hypothesis is correct.

---

[37] Metonymy is the supplantation of an entity's usual name by a phrase or object associated with it - e.g. "the Whitehouse" being used to refer to the US Presidency. In this example, it is debatable whether the Whitehouse relates to a location, a person or an organisation.

[38] Levy, O., Zesch, T., Dagan, I. and Gurevych, I., 2013, August. Recognizing Partial Textual Entailment. In ACL (2) (pp. 451-455).

[39] Levy, O., Dagan, I. and Ramat-Gan, I., 2016, August. Annotating relation inference in context via question answering. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Vol. 2).

3.5. Dataset Size

The level of ambiguity in the dataset is one factor that will influence the decision about corpus size. In general terms, determining the optimal size of a gold standard dataset involves making a trade-off between the value of additional data, compared to the cost of their generation. If the gold standard dataset is intended for the training of text processing algorithms, then the value of increasing its size lies in enabling either more-discriminating classification performance (e.g. higher precision and recall), the modelling of more text features, or some combination of the two.

The question of determining optimal (or alternatively minimum) sample size is a well-studied problem for statistics in general. However, although the issue has been studied for NLP in some domains, there is no general formula for determining the required dataset size for given values of precision, recall, and number of features. The relationships between these parameters depends on aspects of the text-generating process - such as base frequencies of particular types of entity, the prevalence and diagnosticity of disambiguating text features, or the level of irreducible noise - that will differ between domains and corpora. The comprehensiveness and accuracy of pre-existing gazetteers will also influence the relationship between gold standard dataset size and the resulting performance of an NLP algorithm.

An idealised solution is to use progressive (or 'active learning') sampling, in which the gold standard dataset is steadily increased, in parallel with algorithm training, until marginal performance improvements are outweighed by cost. However, in many cases, as here, this is not feasible for practical, programmatic reasons. Consequently, our approach to determining sample size has been made with reference to empirical studies of similar problems and in similar domains.

Studies of the influence of gold standard dataset size on classification performance indicate that performance levels off relatively quickly, and follows a power-law learning curve[40], suggesting that the information content of annotated text is high, but that any two samples are likely to contain relatively high mutual information. Existing gold standard annotated datasets typically contain hundreds or thousands of sentences (see Wissler et al. (2014)[41] for an example).

In practice, of course, the sample size for this project is limited broadly by the resources available, so the main concern has been to confirm that the resulting dataset is not unreasonably small for its intended purpose. Because of the decision to use cross-referenced, crowdsourced annotations to develop the dataset, there is a trade-off between numbers of unique documents, and numbers of distinct annotators (so that, in general, costs are linear in the product of the two variables). The decision to use of the order of 10,000 instances, each appraised by 5-10 annotators, was made so as to correspond in order-of-magnitude terms with common practice in the field. Given the apparent domain- and source-dependence of NLP performance for any given sample size, however, we will not be in a position to evaluate fully this sample size until the dataset is used to train entity- and relationship-recognition algorithms.

[40] Figueroa, R.L., Zeng-Treitler, Q., Kandula, S. and Ngo, L.H., 2012. Predicting sample size required for classification performance. BMC medical informatics and decision making, 12(1), p.1.

[41] Wissler, L., Almashraee, M., Díaz, D.M. and Paschke, A., 2014. The Gold Standard in Corpus Annotation. In IEEE GSC.

Nevertheless, we are confident that analysis of the performance of NLP algorithms on the dataset will provide insights into the likely benefits of its future expansion. Figueroa et al. (2012)[42] show that the analysis of algorithm performance on very small datasets (of the order of hundreds of documents) within a domain can enable accurate prediction of performance against larger ones. Therefore, the planned corpus size for this project should be of sufficient size to enable a robust evaluation of the benefit of its future expansion. Future expansion of the dataset might relate either to an increase in its size (i.e. increasing the number of documents to multiply the occurrence of instances within given classes of entities or relationships), or the sophistication of the tagging (e.g. creating additional classes and subclasses of entity/relationship or adding coreference between increasing consecutive parts of text and addressing entity linking and disambiguation.

3.6. Dataset Content

In addition to the overall size of the dataset, it is also important to consider its content in terms of the nature and the subject of the documents it contains. Marrero et al. (2012)[43] discuss the need for what they refer to as 'content validity'[44] in the testing of information extraction systems. They argue that an evaluation task should reflect the needs and domain interests of the users it is intended for, otherwise assessed performance can significantly exceed actual performance on a 'real' dataset. By extension, this requires training and validation datasets to be reflective of the datasets dealt with by the intended user.

An intelligence analyst is often required to identify and process documents relating to a particular geopolitical subject. Sometimes they will be required to conduct research on a specific event or individual in order to perform detailed and targeted analysis, on other occasions they must keep a watching brief for ongoing events in a region or given field in order to try to anticipate future developments. For both kinds of tasks an analyst is often presented with a dataset which has been filtered for its relevance to the topic of concern, or alternatively begins their analysis by applying subject relevant filters to a broad dataset (e.g. through a google search).

In this context, optimal relevance is usefully defined according to Wilson and Sperber's relevance theory[45], as a ratio between: 1. How much a document contributes to the reader's goal; and 2. How much effort it takes for the reader to access and process the document.

The first term in this ratio depends on the perception of a reader's goals[46]. For the purposes of this study this aspect of relevance is therefore assessed using a subjective judgement about the extent to

[42] Figueroa, R.L., Zeng-Treitler, Q., Kandula, S. and Ngo, L.H., 2012. Predicting sample size required for classification performance. BMC medical informatics and decision making, 12(1), p.1.

[43] Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J. and Gómez-Berbís, J.M., 2013. Named entity recognition: fallacies, challenges and opportunities. Computer Standards & Interfaces, 35(5), pp.482-489.

[44] It should be noted that this differs slightly from the traditional use of the term as used in the scientific method.

[45] Wilson, D. & Sperber, D. (2002). Relevance theory. University College London Working Papers in Linguistics, 14, 249-287.

[46] Rouet, J.F. and Britt, M.A., 2011. Relevance processes in multiple document comprehension. Text relevance and learning from text, pp.19-52.

which a document's content addresses the 'topic' of the dataset[47]. This judgement has been classified as 'Relevance' in the nomenclature of the present project, and its application to the dataset is discussed in greater detail in section 4.3.

The consequent term in the ratio relates to the ease with which the reader can access a document's information (i.e. the level of processing effort required to extract the relevant information). This is partly a feature of the concentration of task-specific information within the document. This has been captured in this project through the development of the construct of 'Richness', which is defined here as the density of entity instances within a text, and is measured as the average number of words (tokens) occurring per entity throughout the entire document. The application of the Richness measure is also further described in section 4.3.

In order to replicate the variation in a typical intelligence analyst's dataset that inevitably results from imperfect querying and filtering, the documents in the dataset will be drawn from different sources (media, government and unaffiliated information producers) and represent a range of Relevance and Richness scores. Supplying a selection of different types of documents with different characteristics could also contribute additional learning and evaluative qualities to the dataset. For example, various studies have shown the impact of instance density (or Richness in the terminology used in this study) on the performance of information extraction systems[48].

Therefore, having a mix of types of writing with a variety of Relevance and Richness within the gold standard dataset, should pose a more realistic and diverse set of challenges to any automated extraction tools which are trained or tested using the dataset. Further, stratifying the dataset based on Relevance and Richness would allow automated extraction performance to be evaluated in relationship to different subsamples of the dataset (e.g. performance against documents with high Relevance and high Richness in comparison with performance against documents with low Relevance and low Richness). This could provide a more sophisticated measure of performance.

There are additional considerations relating to the content of the dataset that are unrelated to its use as a training or evaluation set. These factors stem from the fact that the dataset is being developed for government use, and that there are therefore a number of legislative and policy restrictions which must be taken into account. For example, the compilation of the dataset will involve storing and potentially re-publishing material written by organisations and individuals outside of government. There is therefore a requirement to adhere to UK copyright law[49] and to consider international copyright legislation with regards to content produced under the protection of other national jurisdictions.

[47] In consultation with Dstl it was decided the topic of the gold standard dataset compiled for this project would be the current conflict in Syria and Iraq.

[48] For example, Murphy, T., McIntosh, T. and Curran, J.R., 2006, November. Named entity recognition for astronomy literature. In Proceedings of the Australasian Language Technology Workshop (pp. 59-66) and Ljubešić, N., Stupar, M., Jurić, T. and Agić, Ž., 2013. Combining available datasets for building named entity recognition models of Croatian and Slovene. Slovenščina 2.0: empirical, applied and interdisciplinary research, 2, pp.35-57.

[49] See the following URL for an index of relevant documents: https://www.gov.uk/topic/intellectual-property/copyright

Additionally, some information considered for inclusion in the dataset has the potential to contain personal information about individuals. Understanding the implications for the collection of such data of legislation such as the Regulation of Investigatory Powers Act 2000[50], the Investigatory Powers Act 2016[51] and the Data Protection Act 1998[52] will be essential for making decisions about what should and should not be included in the dataset. In light of the legal nuance involved in interpreting whether and how these acts apply to the dataset, the decision was taken to avoid the inclusion of any material which might constitute personal information and so be subject to this legislation.

Similarly, given the intention to publish the dataset, due consideration was also given to the body of law covering libel, in order to avoid the inadvertent publication of libellous or otherwise defamatory material. Finally, given the selected topic of the dataset (the conflict in Syria and Iraq) the potential exists for source documents containing graphic imagery to be contained in the dataset. Although only text will be analysed within the project, links back to the original source material will be included in the dataset. To reduce the risk of exposing researchers to graphic or disturbing content, any documents which might be considered to fit this category will be omitted from the dataset.

For a full description of how sources and documents are triaged using this set of restrictions within the DSRF process see sections 4.2 and 4.3.

3.7. Summary

The discussion of existing research, evidence and theory laid out in section 3 of this document, in combination with the practical considerations and experiences arising from the development of the dataset, are the foundation for the following set of DSRF parameters which help to define the gold standard dataset for this project:

- The dataset will comprise task-specific documents focussed on the topic of the conflict in Syria and Iraq;
- The dataset will include a range of source and document types, which will be assessed for their 'Relevance' and 'Richness' in line with the DSRF methodology;
- Efforts will be made to minimise the risk that the dataset will include information which might contravene copyright legislation, be subject to data privacy/investigatory legislation, or contain graphic or potentially libellous material;
- The dataset will be tagged using the Baleen entity schema and a relationship schema which has been designed to be task-specific and contain the minimum number of useful relationships;
- The dataset will be annotated at the sentence level. Coreference and relationship tagging will therefore not extend beyond single sentences within documents;
- Ambiguity relating to the tagging of instances will be recorded in the dataset. The dataset will be of sufficient size to enable an assessment of the value of its future expansion.

[50] Regulation of Investigatory Powers Act 2000

[51] Investigatory Powers Act 2016

[52] Data Protection Act 1998

The following section will describe the application of the DSRF methodology in order to illustrate how the parameters above will be implemented.

4. DSRF Methodology

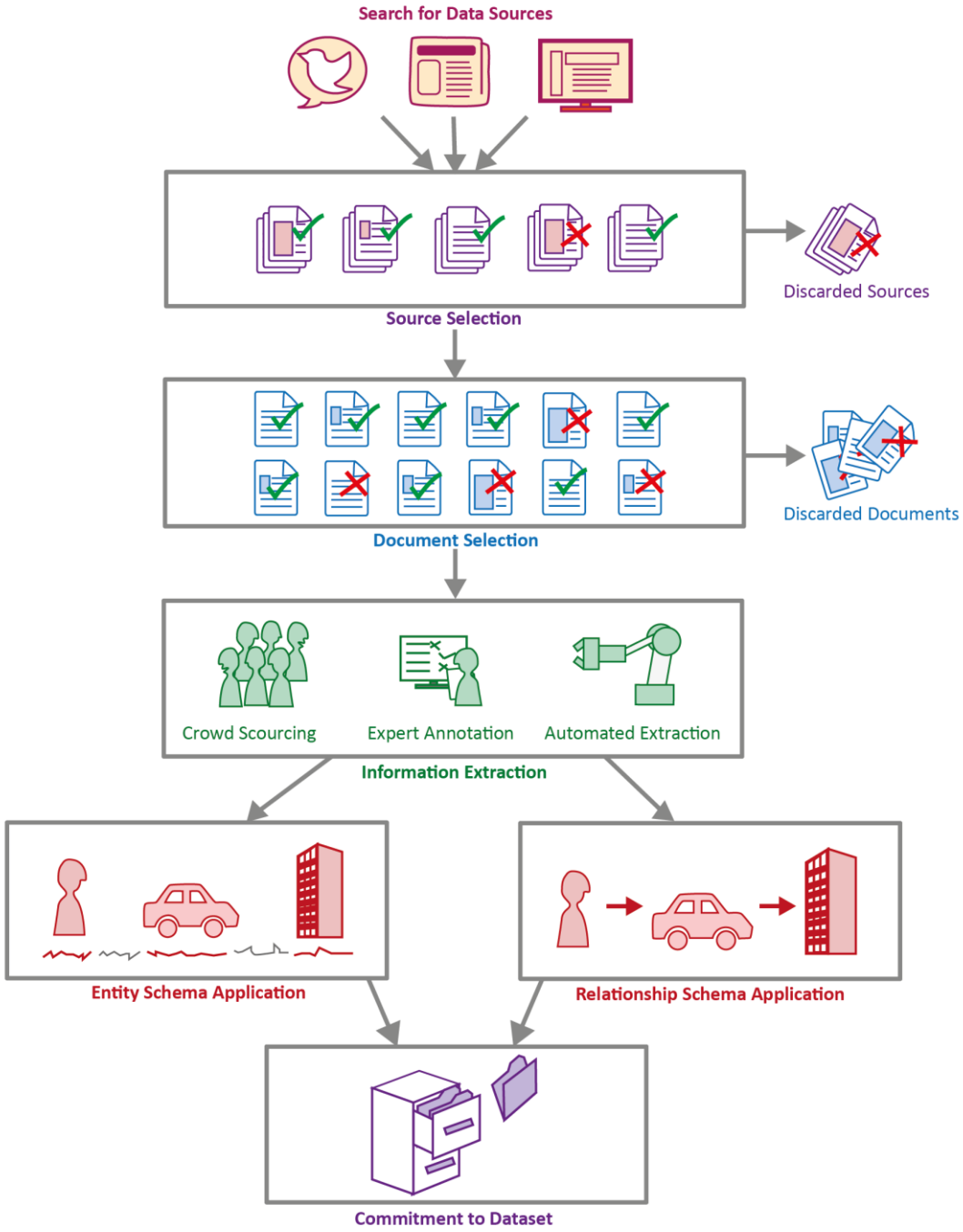


Figure 1 – DSRF Process

The implementation of the DSRF will require a number of steps which progress from the selection of appropriate sources and documents for the dataset, the application of the entity and relationship

schema, and the commitment of the tagged data to the gold standard dataset. Figure 1 shows the phasing of this overall process.

4.1. Search for Data Sources

The first component of the DSRF methodology relates to identifying sources for potential inclusion in the dataset. In this context a **source** is defined as the original producer of a document or set of documents. A **document**, in contrast, refers to an individual piece of content, such as a news article, a tweet or a blog post.

A source might represent a particular media outlet, organisation or an individual. However, in practical terms, it will be taken to mean a URL associated with a given website, a social media account or a particular blog.

Before sources could be assessed they first had to be identified and collected. Initially, potential sources for documents were identified in three ways:

1. Using the study team's existing subject matter knowledge to identify sources of information that provide material which is relevant to the subject of the dataset.
2. Using targeted searches (using search engines) utilising key words associated with the subject of the dataset.
3. Using social media harvesting tools (e.g. Twitter Archiver) to collect tweets which reference key terms related to the subject of the dataset.

This data collection process was targeted to gather information that pertains to the conflict in Syria and Iraq. It should also be noted that for the dataset, only sources producing material in English were considered for inclusion[53]. Initially, fewer than 100 sources were collected in order to minimise the effort required to vet different sources. Should any additional sources be required at a later stage in the project, this step can be reapplied.

4.2. Source Selection

Once collected, sources were then assessed for the following criteria[54].

- Copyright restrictions;
- Personal Data;
- Graphic Content;
- Potentially Libellous Material.

[53] I.e. sources which produce documents written in English (where the vast majority of the text is English language), thus removing the added complication of considering translation of natural language. The collection process employed factored this requirement into the search parameters.

[54] It should be noted that the DSRF process involves the recording of information which is pertinent to the decision about whether to include documents in the dataset. This information along with a recommendation regarding the decision, based on the information available, is provided through the DSRF process; however, the ultimate decision to include and publish documents within the gold standard dataset is taken by Dstl, the owner of the dataset upon its completion. A list of sources for inclusion in the dataset (following their triage) was presented to Dstl for their approval.

The application of the overall DSRF methodology (at source and document level) aims to reduce the information risk to Dstl posed by republishing material produced by other organisations in the form of the dataset. Therefore, the general principle adhered to for restrictions within the DSRF process was to err on the side of caution with regards to excluding sources from the dataset (i.e. in the presence of doubt about a source for a given criteria, it was not included).

The following sections of this document provide more detail about the assessment of each of the four criteria. For full details of the information recorded for these criteria, see the Data Sources and Documents Catalogue Spreadsheet attached along with this technical memo.

Copyright:- The consideration of copyright is particularly relevant to this dataset because not only will the information be collected and stored on behalf of Dstl, the intention is also to publish the dataset upon its completion. It is therefore necessary to consider the copyright restrictions that apply to the sources from which the documents will be drawn. As part of the DSRF process, each potential source's copyright restrictions were recorded; however, the methodology does not provide a legal interpretation of the copyright restrictions, merely a record of them as provided by the source itself. Where none are provided, this absence was noted. The primary considerations regarding copyright for each of the sources, related to whether or not (according to the copyright restrictions as presented) the documents can be held by Dstl and whether or not they can be re-published by Dstl. A full list of the types of copyright restrictions encountered and their application to different sources is recorded in the Data Sources and Documents Catalogue Spreadsheet.

Copyright was assessed at the source level (e.g. the platform or site) because copyright is usually defined to cover all content on a given platform or site. In some rare cases, extra copyright restrictions are applied at a document level; where this was detected, these additional restrictions would be catalogued alongside the document or lead to its exclusion from the dataset. Additionally, for some platforms (e.g. social media and open contribution sites) contributors, by agreeing to the terms of use, are agreeing to the copyright restrictions stipulated by the platform provider. These platform level restrictions were also considered as part of the source assessment.

Personal Data:- As discussed in section 3.6, the potential inclusion of personal data in the dataset has a number of implications related to legislation covering data privacy. The aim is therefore to restrict the amount of personal information contained within the dataset. With this in mind, sources which routinely produced documents containing the names of private individuals, their telephone numbers, email addresses, actual addresses and other private information were not included in the dataset. Individual documents were also checked for this type of content at the next stage of document triage.

Graphic Content:- Given the subject matter for the dataset, it was anticipated that some of the sources could include graphically violent and shocking material as part of their content. Any sources which were discovered to contain such content will be excluded from the dataset, to limit the risk of exposure to individuals involved in the research. The dataset itself comprises purely text files; however, URL links are retained within the dataset which will link back to the original content. This could include imagery or video of a sexually or violently graphic nature. Most web-based sources will be drawn from reputable media sources, thus limiting the risk of exposure. For social media, the content was viewed in purely text format first, enabling the reviewer to assess the risk posed by the full content (including any attached imagery).

Potentially Libellous Material:- The implications of libellous or otherwise defamatory information being contained within documents were considered during the source selection process, in light of Dstl's plan to publish the dataset. An added complication relates to content which was not considered libellous at the time of original publication, but which might subsequently be ruled otherwise. Such material would still reside within the published dataset, even if the original content had been withdrawn by the original producer. To mitigate the risk of the published dataset containing libellous material, sources which were found to contain potentially defamatory information, or specific contentious statements or allegations relating to individuals were not included in the dataset.

4.3. Document Selection

Once sources that were not ruled out by the restrictions from the previous step were identified, the individual documents they produce were then to be examined. These documents are checked to ensure that their content doesn't contravene any of the restrictions applied at the source level (i.e. checked for copyright, personal data, graphic material and potentially libellous material). It is therefore possible that individual documents produced by a source may be excluded from the dataset, while others from the same source are included. However, the primary purpose of the document level assessment is to assess documents for their 'Relevance' and 'Richness'.

Relevance:- As discussed in section 3.6, Relevance refers to how closely a document relates to the overall subject of the dataset. The aim for the dataset is to assemble documents with a range of levels of Relevance in order to match the mixed datasets which intelligence analysts must deal with. Following an initial review, each document was accorded a Relevance assessment. The reviewer provided a subjective assessment using the categories below:

- High Relevance - Primarily about the conflict in Syria/Iraq
- Medium Relevance - Partially relates to the conflict in Syria/Iraq
- Low Relevance - Peripherally mentions the conflict in Syria/Iraq

An automated solution for assessing Relevance was considered (based upon incidence of various key phrases within the text), but it was concluded that key word frequency represents a poor correlate of Relevance as defined within this project (see section 3.6). Given the purpose of the Relevance score was to ensure a mix of content, a subjective score applied by a human reviewer was considered to be most appropriate form of assessment.

Richness:- Richness was designed as a measure of the concentration of entities (and therefore presumed relationships) within a document. As covered in section 3.6, varying the Richness of the documents within a dataset may have an effect on the ability of automated extraction tools to detect, or learn to detect, entities and relationships. It was therefore decided to measure document Richness and capture a range of Richness within the dataset. As such, Richness is applied as a score across a whole document and is measured as the average number of words (tokens) occurring per entity throughout the entire document (as detected by Baleen's automated extraction). What constitutes high, medium or low richness will be measured relatively across the dataset once all the documents have been entered and Richness scores will be updated following expert annotation.

The document level assessments for each document are recorded within the Data Sources and Documents Catalogue Spreadsheet attached with this report. This is a working document and is subject to updating as new documents are constantly added to the dataset. Until the dataset is completed the entries for documents and associated judgements should be regarded as a snapshot of part of the dataset.

4.4. Information Extraction

As documents are selected for the dataset, they will then be processed through the information extraction pipeline developed for this project. This will in effect lead to the tagging of the dataset, where entity and relationship instances will be annotated within the documents according to the task-specific schema described in sections 4.5 and 4.6. The full method for this process will be described in the final technical report delivered for this project; however, in general terms it consists of a hybrid approach, which combines automated extraction using Baleen, expert annotation using a bespoke tool developed for Lot 2 of the Dstl 2016/17 Data Analytics project, and crowdsourced tagging and verification through Amazon's Mechanical Turk platform[55].

4.5. Entity Schema Application

The entity schema for Baleen was developed prior to this project and was therefore a prescribed element of the DSRF and the resulting extraction process. However, there are a number of areas where aspects of the schema must be interpreted, such as the definitions of the entity categories (which are shown in Table 1). The schema itself represents a two level hierarchy, which comprises overall category types (hypernyms) and individual examples of entities which fit in the category (hyponyms). For example, in the category type "vehicle", the following examples would all represent specific examples of a vehicle: car, aeroplane, Toyota, Boeing 747, Air Force One, John's boat. This is in contrast to more complex nested hierarchies which might include multiple levels of hypernyms and hyponyms (e.g. vehicle - class of vehicle (car, boat, aeroplane, etc.) - types of given vehicle (saloon car, speedboat, biplane) - particular models of types of given vehicle (Ford Focus, Aerobus A380, etc) - specific instances of a vehicle (the car with a particular registration number, the boat belonging to a given individual, etc.).

The simpler style of schema employed by Baleen means that anything fitting the overarching description of the category type will be tagged as belonging to that category. However, even this broad use of categorisations still requires an interpretive definition in order to reduce ambiguity arising not from the text itself, but from the interpretation of the schema. Fort et al. (2009)[56] highlight the need for entity definitions which guide annotators, but are not too prescriptive, giving the opportunity for annotators to use their judgement. The table below shows the descriptions that were produced for annotators and crowd workers to guide their judgements alongside the category type as stipulated by the Baleen entity schema.

[55] Follow URL for details of the platform: https://www.mturk.com/mturk/welcome

[56] Fort, K., Ehrmann, M. and Nazarenko, A., 2009, August. Towards a methodology for named entities annotation. In Proceedings of the Third Linguistic Annotation Workshop (pp. 142-145). Association for Computational Linguistics.

Category Type	Description
CommsIdentifier	An email address, a telephone number, a twitter handle or some other social media account
DocumentReference	A number or other identifier which might be used to identify a document
Frequency	A radio frequency
Location	A location or a place of some description
MilitaryPlatform	A military vehicle, ship or aeroplane or other platform to which weapon systems can be attached
Money	An amount of money or reference to a currency
Nationality	A description of nationality, religious or ethnic identity
Organisation	A group of people, a family, a nation state, a government, a business or another type of organisation
Person	A single person
Quantity	A quantity or an amount of something
DateTime	A date or a time
Url	A URL or web address
Vehicle	A non-military vehicle, ship or aeroplane
Weapon	A weapon of some kind

Table 1 - Entity Schema

Even with this set of definitions and allowing for annotator interpretation, there are occasions where specific uses of language will require consistent application of the schema by the expert annotators.

For example, in the sentence - "The Prime Minister met with the leaders of Saudi Arabia, Bahrain, Qatar and the United Arab Emirates" - there is a decision to be made about whether the second entity in the sentence (after Prime Minister) is a set of single persons (e.g. the individual heads of state of Saudi Arabia, Bahrain etc); an alternative set of single persons (e.g. various important officials from each of those countries); a mixture of the previous two; or an organisation comprising a group collective (all of the leaders of Saudi Arabia, Bahrain, Qatar and the United Arab Emirates together as a group). In these cases the annotators will seek to select the interpretation which is deemed most useful for the hypothetical end user, and which seems truest to the intention of the schema. These interpretations will be captured and shared by annotators when they arise during expert annotation using a Rules Capture Spreadsheet to ensure that they are uniformly applied.

4.6. Relationship Schema Application

The relationship schema developed for this project (see able 2 overleaf) was designed to be as simple as possible[57], while being useful with regards to describing important relationships between the entity types described in the inherited Baleen entity schema. It was designed with a nested hierarchy to match the Baleen entity schema. In this case the levels of the hierarchy are 'relationship class' (which determines what kind relationship quality is being assessed) and 'relationship type' (which provides the possible options for that relationship class). There are clearly more relationship classes that could have been created in the schema, but also more relationship types could have been created per relationship class.

As previously mentioned, relationships will only be tracked within sentences in order to prevent the significant increase in task complexity that occurs with coreference across a whole document. Additionally, although multiple entities might be present within a single sentence, the relationships between them will be tagged on a pairwise basis, meaning that multiple relationships might exist within a single sentence (e.g. Entity A may have a relationship with Entity B, and a separate relationship with Entity C). Similarly, within this schema it is possible for more than one relationship type to apply between the same two entities (e.g. where a sentence shows that Entity A 'communicated with' Entity B, and also that Entity A 'likes' Entity B).

A deliberate decision was taken not to include temporal relationships, because 'events', which typically have a relationship with time, were not part of the Baleen entity schema. For an event to be captured within this schema a relationship, such as a meeting between two entities, would first have to be tagged. Only then could a temporal relationship be attributed to this initial relationship. Creating relationships about other relationships was considered, in consultation with Dstl, to represent a level of abstraction which would lead to an annotation process too complex to implement. A full discussion of the reasons for minimising the schema was presented in section 3.2.

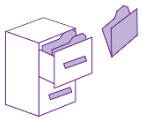
[57] In addition to the problems highlighted in section 3.2 related to a proliferation of relationship types, there are other practical factors to consider when annotating documents. A considerably higher number of relationship instances would mean fewer crowd workers would be able to check each relationship instance. Further, a more complex set of relationship categories would be difficult to explain to crowd workers and could lead to confusion about the intended meaning of stated relationships. It could also lead to relatively few examples of particular relationship types within the dataset, potentially reducing its learning value.

Relationship Class	Relationship Type	Definition
Location The potential for different entities to be proximal in physical space	Is co-located with	Two or more entities in the same place at a given time (e.g. two vehicles being co-located)
	Is apart from	Two or more entities in different places at a given time (e.g. two people being apart)
Ownership The property of entities to be subordinated to other entities, either in terms of ownership or control over	Belongs to	One entity being owned by another entity (e.g. one vehicle belonging to a person)
	Is in charge of	One entity controlling or having responsibility for another entity (e.g. an organisation being in charge of a building)
Metaphysical Connection The capacity for some entities to be attributes of other entities	Has the attribute of	One entity being an attribute or associated quality of another entity (e.g. money having the attribute of a quantity, or a person having the attribute of a nationality)
	Is the same as	One entity being used to mean the same thing as another entity (e.g. multiple names for the same person)
Sentiment The ability for one entity to feel positively or negatively towards another entity	Likes	One entity being positively disposed towards another (e.g. a person likes a military platform)
	Dislikes	One entity being negatively disposed towards another (e.g. a person dislikes another person)
Military Interaction The capability for one entity to be involved in military activity directed towards another entity	Is fighting against	One entity being in armed conflict against another entity (e.g. an organisation is fighting against another organisation)
	Is a military ally of	One entity fighting alongside another entity on the same side of a conflict (e.g. one person being a military ally of another person)
Interaction and Communication The capacity for one entity to contact and communicate with another entity	Communicated with	An entity has having physical contact (i.e. met) or remotely communicating with another entity (e.g. one person communicated with another person)

Table 2 - Relationship Schema

The above relationship schema was constructed to allow the tagging of useful possible relationships between sets of entities in the Baleen entity schema. However, not all of the relationships are possible between all of the entity types. As part of the consideration of the interaction between the entity schema and the relationship schema, a logic meta-model was developed which described all possible relationship permutations for given entity type pairings. For example, a person may 'like' another person, but a vehicle cannot 'like' a weapon. Reducing the number of logically possible relationships in this way, enables a more efficient processing of relationship annotation tasks.

4.7. Commitment to Dataset



Following the annotation of the dataset by Baleen, the expert annotators and the crowd, a calculation of the confidence in the tagging of each instance will be performed as per the principles outlined in sections 3.4 and 3.5 using a variation of the Bayesian Classifier Combination with Words (BCCWords) approach described by Simpson et al. (2015)[58]. The data will be delivered in a MongoDB, with logically structured 'collections' which capture various aspects of the dataset. This will include:

- Details of the data sources and documents, including the assessments of Relevance and Richness;
- Details of the individual source documents, including meta-data (e.g. original URLs where practically possible);
- The complete raw text of each document;
- Tagging of instances of entities and relationships, based on assessments made by the crowd and expert annotators.

The full schema of this database and its collections will be documented on delivery to facilitate use. The task of producing the tagged data in accordance with the DSRF process outlined in this document will form the next part of this project. The execution of the DSRF process will, therefore, lead to the completion of the gold standard dataset.

[58] Simpson, E.D., Venanzi, M., Reece, S., Kohli, P., Guiver, J., Roberts, S.J. and Jennings, N.R., 2015, May. Language understanding in the wild: Combining crowdsourcing and machine learning. In Proceedings of the 24th International Conference on World Wide Web (pp. 992-1002). ACM.

GLOSSARY

Ambiguity - The level of uncertainty in relation to the meaning of a given text segment. In practical terms this is measured as the extent of disagreement between different sources of annotation.

Annotation - The process of an individual or extraction tool ascribing an entity or relationship type to a given item of text.

Document - A complete text document that forms a discrete piece of work, such as a news article, blog post or individual report.

DSRF - The dataset requirement framework, which forms the basis for the content and the structure of this gold standard dataset. It comprises the set of criteria used to select data sources and documents for inclusion within the dataset and the schema used to tag both entities and relationships in the text of the documents included in the dataset.

Entity - Usually taken to refer to the proper names within a given text, which are extracted in Named Entity Recognition (NER). Entity categories within tagging schema might include persons, places and organisations, but can also include other kinds of specific nouns (e.g. vehicles); however, some schema include other types of properties (e.g. quantities or times).

Gold Standard - There is no universally accepted definition of a gold standard dataset for NLP, but it is generally considered to be a dataset which has been extensively tagged by a range of human annotators whose outputs are cross-validated against one another, and against 'objective' automated tagging; inter-annotator agreement is subsequently calculated to ensure quality.

Inference - In the context of information extraction, inference is taken to refer to the process of extracting meaning from text which is not explicitly present in the text itself.

Relationship - The way in which two or more entities are connected or interact with one another.

Relevance - A subjective measure of how closely a document relates to the overall subject of the dataset.

Richness - The density of entity instances (and therefore presumed relationship instances) within a text. It is measured as the average number of words (tokens) occurring per entity throughout the entire document.

Schema - A taxonomical framework for classifying concepts (e.g. entities or relationships) within text. It is often accompanied by definitions of each of the outlined categories.

Source - A producer of text documents, which might be an organisation, an individual or a particular platform.

Tagging - A composite of all the data recorded about a given instance within a text, consisting of the combination of all annotation attached to that instance.

Textual Entailment - An approach to aid natural language inference, which involves generating a hypothesis (h) about the meaning of a given piece of text (t) and then assessing the probability of this hypothesis being true.