Information & Security

G. Valiyev, M. Piraino, et al.
vol. 47, no. 2 (2020): 187-202
https://doi.org/10.11610/isij.4713

Published since 1998

ISSN 0861-5160 (print), ISSN 1314-2119 (online)

Research Article

Initial Exploitation of Natural Language Processing Techniques on NATO Strategy and Policies

Giavid Valiyev [a] (✉), *Marcello Piraino,* [b] *Arvid Kok,* [a] *Michael Street,* [a] *Ivana Ilic Mestric,* [a] *Retzius Birger* [a]

[a] *Data Science Team, Service Strategy, NATO C & I Agency, The Hague, Netherlands https://ncia.nato.int*

[b] *NATO Headquarters C3 Staff Brussels, Belgium*

ABSTRACT:

This paper describes initial exploitation of Natural Language Processing (NLP) techniques applied to a specific set of related NATO documents. In particular, the text similarity technique was applied to document sets with the aim of capturing the relationships between documents or sections of documents from semantic and syntactic perspectives. Thesaurus and triple extraction techniques allowed the understanding of the sentences beyond the syntactic structure, thus improving the accuracy in capturing similar content across documents with diverse syntactic structures. The objective is to assess whether Natural Language Processing tools can retrieve relationships and gaps between such kinds of textual data. This work improves interoperability in NATO by enhancing the development and application of policies, directives and other documents, which dictate how Consultation, Command and Control (C3) systems across the Alliance interoperate and support NATO's operational needs.

ARTICLE INFO:

RECEIVED: 08 JUN 2020
REVISED: 14 SEP 2020
ONLINE: 22 SEP 2020

KEYWORDS:

data science, NLP, text similarity, semantic similarity search, triples, thesaurus, machine learning

✉ E-mail: giavid.valiyev@ncia.nato.int

Introduction

The NCI Agency is NATO's technology and cyber hub. The Agency provides C4ISR (Command, Control, Communications, and Computers, Intelligence, Surveillance, and Reconnaissance) technology, including cyber and air and missile defence. The Agency's Service Strategy and Innovation branch were tasked with creating a new Data Science service to support the analysis of NATO's data, capitalizing the value of NATO's data holdings, bringing value to the decision makers across the Alliance and enhancing their decision support tools. The Data Science team provides a wide spectrum of services to NATO, including identification, sourcing and exploitation of cutting-edge and disruptive data science technologies. The team's data scientists are involved in a continuously rolling portfolio of projects leveraging big data analytics, machine learning, artificial intelligence and visual analytics.

Overarching guidance for the development, delivery and interoperability of communication and information systems (CIS) within the NATO Alliance is governed by the Alliance Consultation, Command and Control (C3) Strategy and Policy. The Alliance C3 Policy is an overarching document, developed in three distinct iterations since 2012, and is composed of 13 individual policies and a glossary; all compiled under one reference document.[1] Policies are developed with considerable input from subject matter experts from all nations of the Alliance, with specialist knowledge across the broad scope of military C3. This range of inputs creates the potential for duplication of content, or even of divergent guidance: a situation which is monitored and prevented by the careful stewardship by the NATO HQ C3 Staff.

In the following paragraphs, the authors will discuss the analysis of the relationship between these two documents sets by applying NLP techniques. In particular, Section 2 defines the problem and challenges of subject matter experts. Section 3 summarizes text similarity techniques and compares similarity measures. Section 4 provides an overview of the previous application of NLP techniques. Section 5 explores our dataset and understands its distribution. Section 6 introduces the tools that have been adopted. Section 7 outlines the methodologies that have been implemented. Section 8 presents, evaluates and discusses the result obtained. Finally, Section IX provides conclusions and an outlook on future work.

Problem Definition

The objective of our project is to investigate the use of NLP as a vehicle for supporting the current subject matter expertise, in the analysis and updates of the Alliance C3 Policy. Applying appropriate data science tools and techniques improves the ability to identify and address policy gaps and overlaps at relevant stages. Improving the relationship between the Alliance C3 Strategy and its supporting policy documents, and between different policies themselves, also supports the work of developing lower-level implementation directives.

NLP tools, and text similarity in particular, were applied to the document set of strategy and policies in order to explore how this technology can augment the

current development and stewardship of these essential documents for military CIS. Namely, can text similarity:

- Identify how existing policy documents are aligned to the scope of the Alliance C3 Strategy, with the aim of understanding whether new policy is required for the implementation of the strategy itself?
- Identify the relationships between policy documents in order to address overlaps, gaps, as well as areas of strength (that is, mutually reinforcing statements situated in different policy documents)?
- Identify *concepts*, with the aim of understanding if a particular concept of interest is covered in these documents? Four type of concepts were identified as objects of this work in accordance with the policy authors, namely: Information, Architecture, Capability and Governance.

Rationale for text similarity

Natural Language Processing (NLP) refers to a discipline of data science and linguistics that combines machine learning (ML) and computer science technologies to draw meaning from unstructured text documents. NLP provides a set of toolkits such as Text Similarity, Named Entity Recognition (NER), Text Classification, Sentiment Analysis and many others with an objective of understanding human language and analysing text in order to provide solutions. In recent years there have been big improvements in capturing the context in which text is written; first using word embedding models such as Word2Vec and then through Transfer learning models with attention mechanisms [2] such as Google's open sourced BERT [3] model.

Text similarity is one of the most widely exploited techniques in NLP. The objective in applying text similarity is to understand how close two text strings are to each other based on their syntactic structure and—if possible—semantic meaning. This technique has multiple use cases, such as finding similar documents based on their content. It is implemented in search engines with the aim of retrieving the most relevant documents to a certain query. Question answering systems rely heavily on text similarity as they use this technique in order to understand user queries and retrieve the most relevant answers.

Text similarity measures are often combined with word embeddings, which are a distributed word representation based on a neural network. Distributed word representation embeds every word into a low dimensional continuous space, capturing in this way the semantic and syntactic information and groups similar words. However, training a neural network for word embedding in a domain specific language requires a large training dataset which is not in the scope of this project due to project limitations, but which will be explored in future work.

Similarity Measures

Similarity measures are often used in solving pattern recognition problems such as classification, clustering, information retrievals, taxonomy etc. The objective of such measures is to find a quantitative measure of how far apart two objects are.

In NLP we use such measures in order to measure the text similarity by applying them to documents in the form of vectors where each dimension of vector contains a numeric representation of term.

Within similarity searches in NLP, a few algorithms are typically used: Cosine similarity, Block distance, Euclidean distance, Levenshtein distance, Tanimoto similarity, and Dice's coefficient. Each of these measures is most applicable to different circumstances.

In this case, Cosine Similarity was deemed most appropriate for the scope of this project as it takes into account the length of documents and corrects the difference in length between two documents, we will see in the next subsection why this is an important step and the implications of Zipf's [4] law on such measure. It also considers the frequency of terms in each document, unlike some of other measures such as Euclidean distance.

The difference between Cosine and Euclidean distance is that while the first measure is looking to the distance of angles between two vectors, the second measure is actually measuring the distance and not angle between them. This implies that if we would compare two documents, one much larger than the other, Euclidean distance would detect a large difference between the two just because of their length, without taking into account the weights of terms. Cosine is usually used in cases where the magnitude of vectors does not matter and this is the case when working with textual data represented by some weight.[5]

Cosine Similarity

Cosine similarity is a measure of similarity between two vectors and can be applied to vectors in any number of dimensions. It is widely used in information retrieval and text mining, where each dimension represents a different term and a document is represented by a vector, with the value in each dimension corresponding to the number of times the term appears in the document. Cosine similarity provides a useful measure of how similar two documents are likely to be in terms of their subject matter. Mathematically it measures the cosine of the angle between two vectors (vector representation of documents in our case) projected in a multi-dimensional space.

Related Work

An initial exploitation of NLP techniques to assess the alignment of NATO Policy documents [6] has shown some promising results and possibilities of potential utilization of such solutions in support of staff officers and analysts. To some extent such solutions were applied with the aim of understanding and capturing the overlap between different set of documents, based on concepts covering such documents. Techniques such as Bag of Words (BoW) and counting term co-occurrence provided an initial idea about the concepts being covered in these documents with the aim of understanding the emphasis of two document sets.

However, such experiments also highlighted the limitations of using syntactic approaches which do not capture an understanding or semantic meaning of text.

Therefore, we adopted a hybrid approach which tries to overcome the limitations encountered in previous applications of NLP in our domain with the aim of enhancing syntax-based approaches, with usage of domain specific thesauruses and dependency analysis of structure of sentences.

Dataset and Structure

The initial dataset consists of thirteen documents: the Alliance C3 Strategy (01-AC3S) and twelve policy documents:

- Services Management Policy (02-SMP)
- Lifecycle Management Policy (03-LMP)
- Waveform Policy (04-WP)
- Interoperability Policy (05-IP)
- Federation of Communication Services Policy (06-FCSP)
- Software Policy (07-SP)
- Capabilities Implementation Policy (08-CIP)
- Enterprise Architecture Policy (09-EAP)
- Cloud Computing Policy (10-CCP)
- Green IT Policy (11-GITP)
- IPV6 Policy (12-IPV6P)
- Data Management Policy (13-DMP)

Alliance C3 Strategy and policy documents are highly structured documents with a consistent format. For the purpose of this analysis the principles section of each document has been considered as most valuable, following guidance from policy experts, as they contain the prime aim of each document. These sections list high-level principles expressed in a maximum number of four sentences each. The Alliance C3 strategy consists of 24 principles while policy documents contain between 6 and 15 principles; overall we find 148 individual principles within the policy documents.

Before pre-processing, our vocabulary consists of 583 terms and 1664 tokens. We have analysed the distribution of terms in order to understand if our dataset follows Zipf's law,[4] which states that given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table. Thus the most frequent word will occur approximately twice as often as the second most frequent word, three times as often as the third most frequent word etc. This implies that there is a relationship between the probability of a word occurring in a corpus and its rank (in terms of frequency). The probability function in Zipf's law is defined as:

$$f(k;s,N) = \frac{1/k\text{^}s}{\sum_{n=1}^{N} (1/n\text{^}s)}$$

Where *k* is the rank of term, *N* is the vocabulary size (583) and *s* is the parameter of probability distribution and is set to 1 in classic Zipf's law. Zipf's law implies a Power law [7] relationship between the rank of term and its probability of occurring in dataset.

Table 1. Term distribution in our dataset with proportion.

Rank	*Term*	*Frequency*	*Proportion*	*Predicted_prop.*
1	are	66	0.1132	0.1132
2	is	66	0.1132	0.0566
3	and	60	0.1029	0.0343
4	nato	42	0.0720	0.0180
5	the	41	0.0703	0.0140
6	to	37	0.0634	0.0105
7	c3	27	0.0463	0.0066
8	of	27	0.0463	0.0057
9	software	26	0.0445	0.0049
10	capabilities	24	0.0411	0.0041

We have applied Zipf's probability calculation to our dataset in order to have an initial understanding of its distribution and validate against the view of the domain experts: that policies and strategy documents are written with a specific domain language depending on the topic covered in each document, as shown in scatter plot in Figure 2. Zipf's law applied in NLP indicates that a few terms are used very often and many terms are used very rarely. We can say that most used terms are generally responsible for the majority of overall terms occurring in the entire corpus. Alignment with Zipf's law indicates good accuracy from selecting Cosine as our similarity measure.

Tools

KNIME Analytics Platform

KNIME analytics platform [8] used for much of this work is one of the leading open source data mining tools available; it is language agnostic and provides a graphical composition framework for data preparation, model fitting, and result analysis. KNIME provides a complete set of solutions for NLP tasks with dedicated "text processing" node components which allow the user to build NLP workflows with state-of-the-art solutions.

KNIME provides an easy integration with some other leading tools used in data science such as Python and R when needed. This makes the data scientist's job flexible, allowing the use of the best combination of these tools and possibility to approach the problem from different perspectives, combining different and compatible solutions.

Microsoft Power BI

The NCIA's Data science team has adopted Microsoft Power BI as a visualization tool in recent years. Although KNIME provides the possibility to explore the results

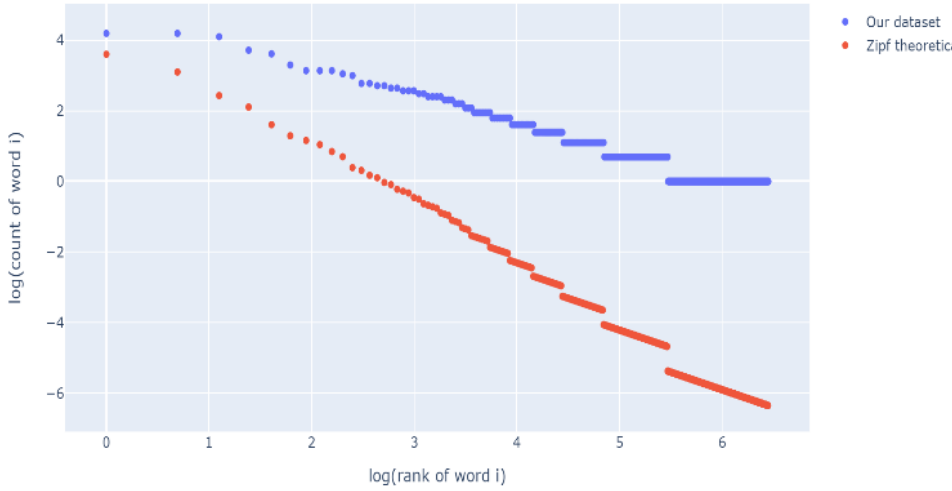

Figure 1: Relationship between the rank of a word and distribution within dataset and Zipf's probability.

by plotting them in specific data visualization nodes, the presentation is not as compelling as it could be with more specialized data visualization tools. Given the availability in NATO domains and ease of use of Microsoft Power BI we have opted to create interactive dashboards allowing domain experts to visually explore and gain insights into the reasons behind the scores.

Methods

Prior to explanation of the complete NLP process, it is important to explain some of the important components of this work and their role in weighting the terms (most important terms carrying much of the information) in our corpus at sentence level. In particular we are referring to:

Glossary

The Glossary is a collection of technical and domain-specific terms. It consists of more than 30,000 terms and definitions. It was developed from many publicly available and NATO specific glossaries such as; the NATO Terminology database (maintained by the NATO Standardization Office (NSO) [9] and containing particular terms and abbreviations used in NATO/military environment); C3 Policy Glossary [1] (provided with the Alliance C3 Policy, containing terms used in the development of the policy with related definitions). It also included terms from publicly available glossaries such as those used of the Information Technology Implementation Library (ITIL),[10] The Open Group Architecture Framework (TOGAF),[11] as well as more generic management glossaries of relevance such as those for *PRINCE2* [12] project management and *Managing Successful Programmes*.

The glossary is used to distinguish known domain specific multi-word terms where possible.

This is an initial step in improving the context in which terms are occurring. The role of multi-words is important in improving the final accuracy of similarity scores.

As an example, let's take two scenarios in order to explain the importance of multi-words:

- "The management cares about employees"
- "Metadata management is mandatory for employees."

In scenario number one we will use glossary in order to tag multi-words in these sentences. The result of this process will lead us to have the term "Metadata management" as a multi-word term. When applying the similarity measure on these two sentences the term management from the first sentence will not be considered an overlap with the second sentence where "management" is considered a multi-word term used in context of "metadata management".

In scenario number two we omit the step of tagging multi-words. When applying similarity measures to these two sentences we will find the term "management" in both of sentences and this will lead to have a high similarity score which will be obviously not accurate.

This example highlights an importance of having such a glossary available and if not available, there are some other NLP techniques which allow the development of such multi-word glossaries, term co-occurrence being one of these techniques.

Thesaurus

The basis for our thesaurus is the open-source WordNet database for English language.[13] WordNet is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations.[14]

For the purpose of this work, we have used the WordNet lexical database in order to extract the list of terms with specific relationship type such as synonyms, hyponyms and enhanced this list with domain-specific synonyms and related terms involving subject matter experts in the definition of such list.

The thesaurus was also used in the definition of concepts of interest. Such sources allowed us to retrieve and understand the coverage of four concepts of interest within sentences and documents not only from a syntactic point of view but also being able to capture such concepts if referred to with synonyms or related terms.

The importance and usage of a thesaurus is highlighted by the following example sentences extracted from Federation of Communication Services Policy:

"Federated network service providers establish instructions for interconnection points."

From a syntactic perspective there is no clear reference to our concepts of interest (information, architecture, capability and governance). This is not true if we check the definition of concepts in the thesaurus where we can find "interconnection point" and "federated network" as related multi-word terms to the synset "Architecture." This example also highlights the importance of having domain-specific knowledge available where possible in order to capture the hidden semantic relationships between terms not available in publicly available lexical databases.

Term Frequency

The classical approach of weighting terms (in our case single and multi-word terms) in the document or sentence corpus is by counting the relative frequency of each term (TF) multiplied by Inverse Document/sentence Frequency (IDF). This weight will give an initial idea about the importance of a term given its frequency in our corpus. It is particularly important to apply the IDF value in document sets having different lengths. The underlying logic is that terms repeated too often in the corpus will carry less valuable information compared to the ones with low frequency. This links back to the application of Zipf's law to our dataset.

As an example, most NATO documents are likely to include the term "NATO". To this end, we introduce a mechanism (IDF) for attenuating the effect of terms that occur too often in the document set to be meaningful for relevance determination. We need to scale down term weights of terms with high frequency (TF) and reduce it by a factor that grows with its frequency in other documents where the same term occurs. Our final TFIDF score will give us an initial normalized quantitative score to each of term in the corpus.

Triples

The Stanford CoreNLP [15] toolkit is an extensible pipeline that provides core natural language analysis. It is one of most widely used NLP toolkits with a broad range of grammatical analysis solutions. It excels in achieving different types of complex NLP tasks such as dependency parsing.

The Open Information Extractor (Open IE) [16] is one of the Stanford CoreNLP modules able to extract the grammatical structure of sentences and identify the triples. Extraction of triples from sentences provide the possibility to determine the core aim/meaning of each sentence.

KNIME allowed us to a use pre-trained Open IE machine learning model facilitating the extraction of triples from sentences. This node extracts entailed clauses which then are reduced to their main statement and split into subject, predicate and object.

The way we have used the triples consists of updating previously calculated TFIDF weights and diminishing these values if the term is detected as not being part of Subject or Object of each sentence; with an underlying logic that terms having such grammatical roles should have a higher importance given the fact that they belong to the core part of sentence.

NLP Process

Once we have a clear idea of the weight assigned to terms, we can proceed with the explanation of the entire NLP process applied in this work. The main objective of the NLP process is to transform the terms contained in the strategy and policy documents into vectors, in order to be able to apply similarity measures to these documents at sentence level. This is done in the following steps:

1. Read Text: to read the content of the documents to be analysed and to split the content of the documents into sections. This allows us to focus particular interest on the Principles section of the documents which describe the prime aim of each document.
2. Tokenize: the objective in this step is to transform simple text into a more complex Document object. The tokenization operation identifies and labels parts of the documents as sections.
3. Enrichment: where specific tags are attached to terms. In particular, the Part of Speech tagger node is applied in this step with an objective of defining the grammar roles for terms. Additionally, multi-words are tagged with the help of previously defined glossary.
4. Pre-processing: this step simplifies the content of the documents by removing information which does not add any value. Multiple nodes are applied in this to remove punctuation, filter out numbers and stop words, Case Converter and Stanford Lemmatizer (Lemmatizes terms contained in the input documents with the Stanford Core NLP library to remove inflections of terms). Importance of terms are calculated with TFIDF calculation and normalized after triples extraction.
5. Bag Of Words with related weight assigned and calculated with previously described process: This is a technique used in natural language processing and information retrieval to disaggregate input text into terms. Documents composed of a similar bag of words are supposed to be similar in content. Terms are compared with thesaurus in order to retrieve synonyms and related terms.
6. Similarity search: taking each row in the query table (for strategies and concepts) and searches the reference table (policies) for a number of rows matching the specified similarity/distance criteria (cosine similarity in this case). If multiple results are requested, the query result row is duplicated for each subsequent match.

Improvement

After the application of the NLP process, workshops with subject matter experts were held to improve the thesaurus, in order to achieve better results in the retrieval of concepts in future iterations and better identify the domain-specific synonyms and related terms.

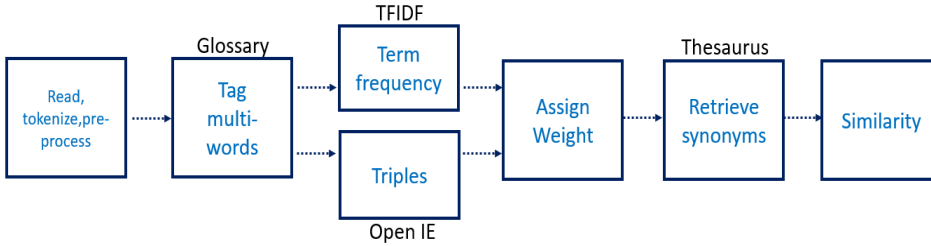

Figure 2: NLP process applied in our work.

Results

The results obtained from the application of cosine similarity measures to strategy and policies provided us with a list of sentences and similarity scores between them. Sentence couples having the highest similarity scores are the ones having the strongest relationships.

We have aggregated such sentences according to the sources (documents) they were extracted from in order to be able to compare results at document and principle level and try to answer our original questions:

- Identification of concepts. Can text similarity identify concepts, with the aim of understanding if a particular concept of interest is covered in these documents?

Four concepts of interest were compared with sentences from policy documents (Figure 4) with the aim of understanding which principles/documents are covering the concepts of interest.

As we can see from the results, the concept "architecture" has the most coverage among the policy documents. In particular, cosine similarity metrics show the Service Management and Enterprise Architecture policies are rich with synonyms and related terms to this concept of interest. This was expected from domain experts as these policies both give direction relating to C3 architecture.

The Data Management policy is a document covering the concept of "information" strongly. This relationship was also expected by the domain experts. Figure 4 also shows the "governance" concept has good coverage in the Service Management policy.

While the concept "capability" does not seem to have a large coverage among the policy documents. This might be due to the definition of the concept "capability" which is more likely to appear in lower level documents which describe specific equipment, training etc.

Identifying Coverage of the C3 Strategy

- Can text similarity identify how existing policy documents cover the scope of the Alliance C3 Strategy; can it "understand" whether new policy is required for the implementation of the strategy itself?

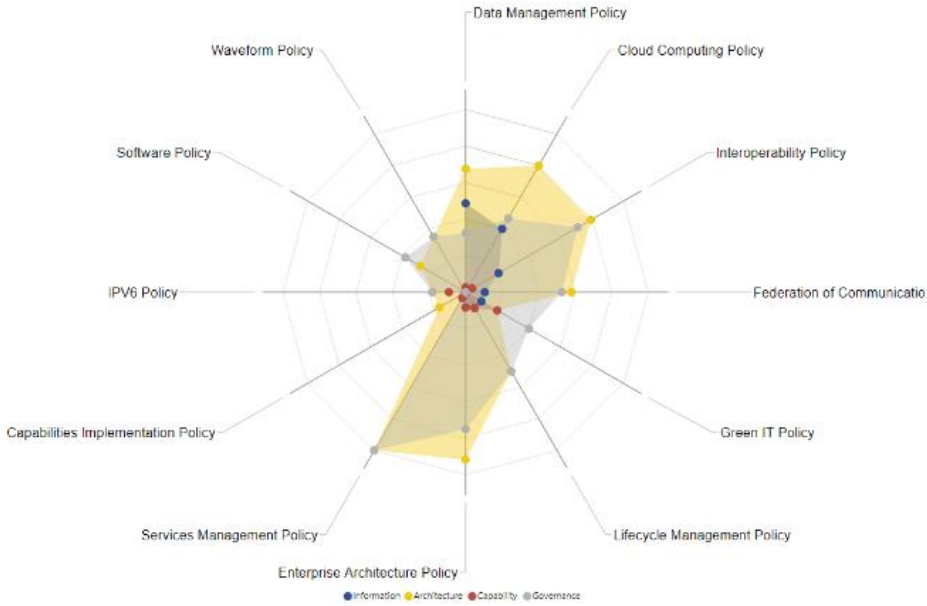

Figure 3: Coverage of concepts within different policies.

Sentences from the Strategy document are compared to sentences from policy documents (Figure 5) with the possibility to compare such documents at individual principle level (one or a few sentences) or at document level (all principles of the same document). The intensity of the blue colour indicates the grade of overlap between the individual principles (rows) from strategy document with principles of policy document (columns).

Identifying Relationships

- Can text similarity identify the relationships between policy documents in order to address overlaps or gaps, as well as areas of strength?

Sentences from policy documents were compared with sentences of other policy documents (Figure 6) with a possibility to compare such documents and identify relationships between individual principles across different policy documents. The results highlighted and confirmed some expectations of the domain experts, such as the limited correlation between Waveform policy and all other policies (outlined in figure 6). The reason for this gap is that the waveform addresses specific aspects of physical interconnectivity which has little interaction with other elements of the C3 domain.

For all of the three questions it was important to be able to explain each of the relationships and display the reasons behind the similarity scores obtained, as shown in Figure 7. Through interactive dashboards it is possible to see how the

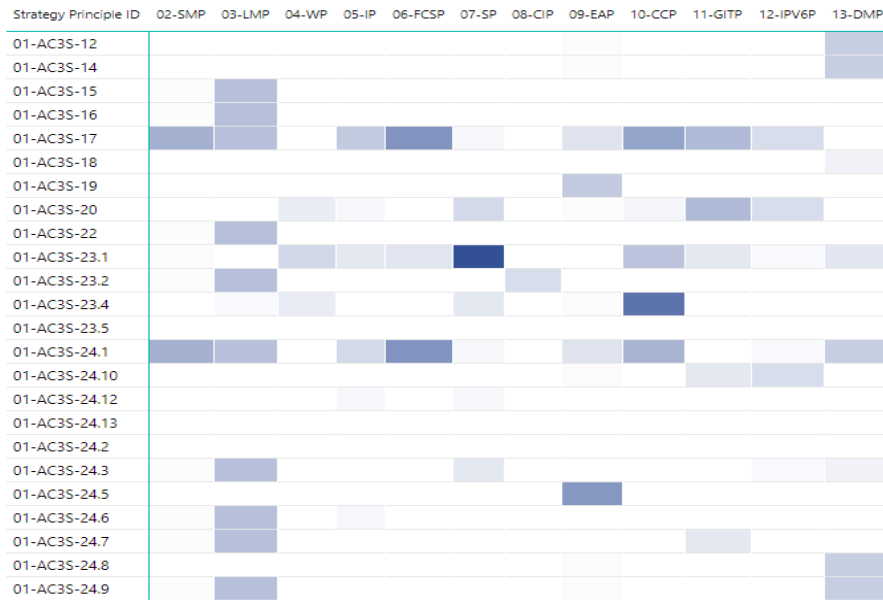

Figure 4: Principles from Strategy (rows) compared with policies.

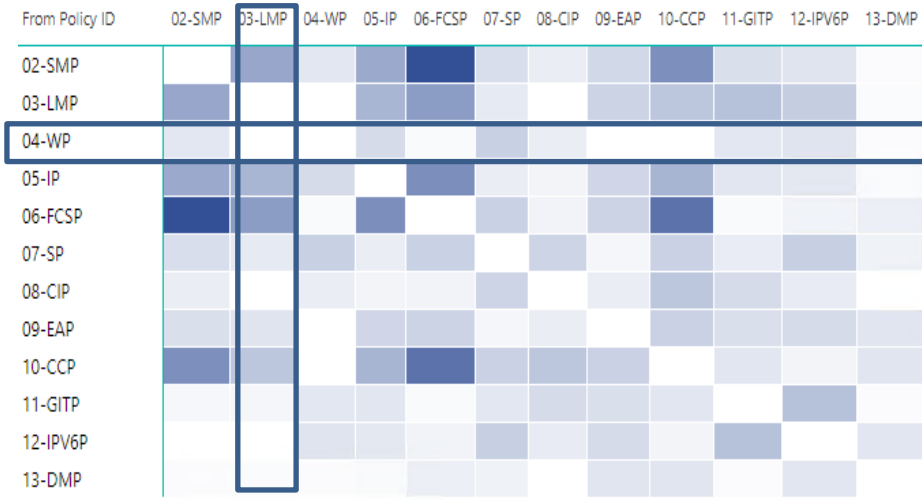

Figure 5: Polices comparison with policies.

workflows created score and weight individual words and triples, in order to identify relationships between the same concepts described in different words across the document set and to “explain” our method.

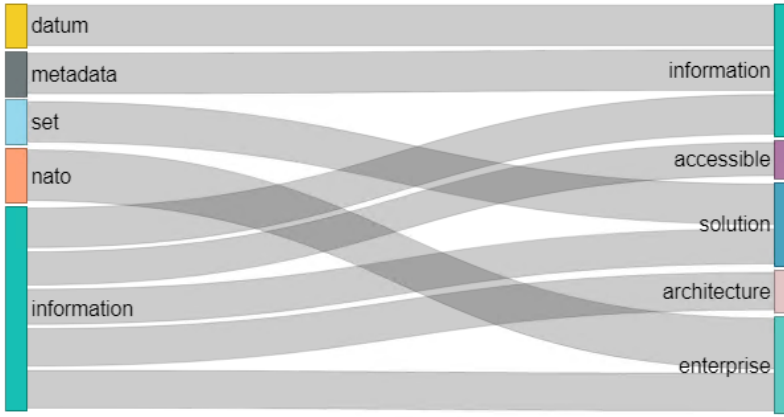

Figure 6: The reason of overlap between two sentences.

When comparing two documents/principles we have explained the resulting similarity scores displaying the overlapping terms (including synonyms and related terms). In Figure 7 we can see an example of the overlapping terms (exact matches, synonyms, and related terms) when comparing two principles.

Conclusion

In a document-rich environment, where many documents provide guidance or direction, NLP has much to offer as a powerful tool to augment the skills and knowledge of experienced domain experts to identify correlations or deviations within document sets. This work deliberately used a small document set, although it is recognized that a larger corpus should provide more accurate results and provide more value, by identifying overlaps, gaps or divergence in the guidance being given in NATO C3 documentation at differing levels of granularity.

With this work we have applied text similarity techniques and gone beyond a simple syntactic similarity. We have assessed the potential value which a domain specific thesaurus can add to NLP results. We have also understood the limitations of such an approach driven by the experience of domain experts in defining domain-specific synonyms and related terms which compose the thesaurus. In future work our aim is to train a Natural Language model on a much larger corpus of NATO-specific documents, including the larger set of C3 directives (which go into greater detail on the same topics) in order to generalize such model between different Communities of Interest (COI) with related document sets.

Initial questions were answered with potential improvements in policy versus policy comparison where after an initial assessment from domain experts, some improvements were suggested in order to enhance the thesaurus and capture hidden relationships.

This work shows that the value of NLP is enhanced significantly when domain expertise is injected into the analysis, through the use of a defined thesaurus. It

also shows the value of using triples to extract and compare concepts found within documents.

Acknowledgment

The authors wish to thank Eleanor Williams, Emilia Dettorres and Ivano Pennacchio whose subject matter expertise and exceptional support made this work possible.

References

1 NATO, “Alliance C3 Strategy,” C-M(2018)0037, 2018.

2 Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, 2017.

3 Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *Proceedings of NAACL-HLT* 2019, Minneapolis, Minnesota, June 2 - June 7, 2019, pp. 4171–4186.

4 Álvaro Corral, Gemma Boleda, and Ramon Ferrer-i-Cancho, “Zipf’s Law for Word Frequencies: Word Forms Versus Lemmas in Long Texts,” *Arxiv*, (2015).

5 Chris Emmery, “Euclidean vs. Cosine Distance,” March 25, 2017, https://cmry.github.io/notes/euclidean-v-cosine.

6 P. Eles, B. Pennell, and M. Richter, “Assessing NATO Policy Alignment,” *2016 International Conference on Military Communications and Information Systems (ICMCIS)*, Brussels, Belgium, 2016.

7 Mark E. J. Newman, “Power Laws, Pareto Distributions and Zipf's Law,” *arXiv:cond-mat/0412004* (2006).

8 Michael Berthold, Nicolas Cebron, Fabian Dill, Thomas Gabriel, Tobias Kötter, Thorsten Meinl, Peter Ohl, Christoph Sieb, Kilian Thiel, and Bernd Wiswedel, “KNIME: The Konstanz Information Miner," in: *Studies in Classification, Data Analysis, and Knowledge Organization* (Springer: 2007).

9 NATO Standardization Office (NSO), https://www.nato.int/cps/en/natohq/topics_124879.htm.

10 "ITIL," *Axelos*, 2020, available at: https://www.axelos.com/best-practice-solutions/itil.

11 "TOGAF," *Opengroup,* 2020, available at: https://www.opengroup.org/togaf.

12 Prince2, https://www.prince2.com/eur.

13 George A. Miller, “WordNet: A Lexical Database for English,” *Communications of the ACM* 38, no. 11 (1995): 39-41, https://doi.org/10.1145/219717.219748.

14 WordNet, https://wordnet.princeton.edu/.

[15] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky, "The Stanford CoreNLP Natural Language Processing Toolkit," *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Baltimore, Maryland, 2014.

[16] Gabor Angeli, Melvin Premkumar, and Christopher D. Manning, "Leveraging Linguistic Structure for Open Domain Information Extraction," *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, China, 2015.*

[17] "C4ISR Architecture Working Group (AWG): Final Report," Department of Defense, Washington D.C., 1998.

[18] David M. W. Powers, "Applications and Explanations of Zipf's Law," in: *New Methods in Language Processing and Computational Natural Language Learning* (ACL, 1998), 151-160.