Domain-Specific Term Extraction: A Case Study on Greek Maritime Legal Texts

Despoina Mouratidis
Ionian University
Department of Informatics
Corfu, Greece
c12mour@ionio.gr

Eirini Mathe
Ionian University
Department of Informatics
Corfu, Greece
c17math@ionio.gr

Yorghos Voutos
Ionian University
Department of Informatics
Corfu, Greece
c16vout@ionio.gr

Klio Stamou
Ionian University
Department of History
Corfu, Greece
h18stam@ionio.gr

Katia Kermanidis
Ionian University
Department of Informatics
Corfu, Greece
kerman@ionio.gr

Phivos Mylonas
Ionian University
Department of Informatics
Corfu, Greece
fmylonas@ionio.gr

Andreas Kanavos
Ionian University
Department of Digital Media and Communication
Kefalonia, Greece
akanavos@ionio.gr

ABSTRACT

Preservation of cultural heritage has significantly attracted many research efforts. Amongst them, significant interest has been presented to sharing of cultural heritage content, whereas one major aspect of cultural content is the one concerning maritime heritage. In order to efficiently consume such content, high-quality descriptions are required. Towards this goal, we propose an approach for automatic term extraction of documents, where our methodology is based on the use of word embeddings along with a deep neural network architecture. We demonstrate its efficiency by using a large corpus of legal documents related with maritime.

KEYWORDS

Deep Learning; Domain-Specific Terms; Gaze Detection; Legal Texts; Maritime Texts; Neural Networks; Term Extraction; Text Tagging

ACM Reference Format:
Despoina Mouratidis, Eirini Mathe, Yorghos Voutos, Klio Stamou, Katia Kermanidis, Phivos Mylonas, and Andreas Kanavos. 2022. Domain-Specific Term Extraction: A Case Study on Greek Maritime Legal Texts. In *12th Hellenic Conference on Artificial Intelligence (SETN 2022), September 7–9, 2022, Corfu, Greece.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3549737.3549751

SETN 2022, September 7–9, 2022, Corfu, Greece

ACM ISBN 978-1-4503-9597-7/22/09…$15.00
https://doi.org/10.1145/3549737.3549751

1 INTRODUCTION

Preservation of cultural heritage that has been inherited from past generations in order to be transmitted to future generations, is significantly important for several reasons such as empowering local communities and enabling several groups of people to fully participate in social and cultural life[1]. With the term "cultural heritage", we mainly refer to artifacts and traditions of a specific community, while by "preservation" we refer to keeping them intact, e.g., by protecting them from physical damage, theft, distortion among generations, etc. Preservation of cultural heritage requires certain actions concerning several research areas. Such procedures include the identification of cultural property, the recording, the storage as well as the archiving of the findings. Other procedures include restoration and/or reconstruction of buildings, sites, etc. More to the point, sharing of cultural heritage content plays an important role to fostering the identity of future generations, while also helps bridging different nations, e.g., through common traditions. Moreover, it may have a serious impact on both areas of tourism as well as economy.

As far as maritime heritage is concerned, it is not solely limited to physical resources, e.g., shipwrecks and sites, but it also includes several types of documents, artifacts and multimedia digital content. Usually, the archives and collections of maritime museums comprise of documents and archives. The former include items of archives along with drawings of ships and maps, while the latter include nautical objects and art. Often the aforementioned collections have been enriched by digital and audiovisual material. More specifically, the main aspects of maritime heritage include among others: a) archive material covering the different aspects of maritime professions and industry (ships, shipyards, headquarters of shipping companies, etc.); b) physical objects and materials such as

[1] https://en.unesco.org/content/preserving-our-heritage

instruments, tools, equipment and maritime-related art; c) libraries that include material related to museums and museology, underwater archaeology, naval history, registries, navigation guides, etc.; d) galleries containing historical photographs of ships, shipwrecks, ship crossings, maritime professions etc.; and e) collections of shipping companies, ship-building and design collections, containing records, text and graphic material.

In order to efficiently manage this content, we should initially process and then model it, so as to be able to produce high-quality descriptions of it. One of the most well-known techniques, which is often used towards this goal, is the automatic term (terminology) extraction [13], which aims to extract relevant terms from a given corpus of documents. This set of terms may in following be used into a plethora of semantic-based applications, such as web-crawlers, web services, web queries [20], recommender systems [8], text mining and natural language processing [21], as well as indexing and retrieval of documents [14], etc. In brief, among the first steps towards modeling a specific domain of knowledge is to define (or in practice extract) a vocabulary of relevant terms. By "relevant" we denote those terms that achieve high relevance in terms of the given domain and are able to describe most of its concepts.

Therefore, in this work we present an approach for the automatic extraction of terms within the broader domain of maritime heritage. Specifically, we first create a dataset consisting of a large corpus of documents. Upon pre-processing, we extract textual features from each document and in following we proceed with a word embeddings approach for representing text data to numeric data. At the next step, a deep neural network architecture is implemented, where its role is to decide whether a given term is a maritime one or not. Since the percentage of maritime terms within the document is significantly smaller than the one of non-maritime terms, our problem is imbalanced and thus, a typical approach may lead to poor performance. To overcome this, we artificially balance the dataset by creating new samples, using an over-sampling approach. We evaluate our proposed methodology using a large and challenging dataset comprising of legal texts on maritime topics in the specific language.

The rest of this paper is organized as follows: Section 2 presents related research works and tools within the broader area of automatic term extraction. In following, Section 3 presents the dataset that has been utilized for the experimental evaluation of this work, the set of statistical features that have been selected to describe a given document, the word embeddings method that has been adopted and the utilized deep network architecture. Experimental results are presented in Section 4, whereas the discussion of the results, the conclusions drawn and the future plans are depicted in Section 5.

2 RELATED WORK

Learning techniques that are chosen to be trained for automatic recognition of conditional words versus non-conditional words include, among others, Decision Trees [3], rule-based techniques [4], Naive Bayes, Support Vector Machines, collaborative learning and deep learning [15]. The recognition performance reaches up to 95% in terms of correct classification rate (accuracy). In our research, we focus on extracting information that can be used to discover terms in the field of maritime heritage described in legal texts. Consequently, we are not concerned to develop a methodology for document analysis in general but rather a textual basis around a specific topic.

2.1 Automatic Term Extraction Techniques

Automatic term extraction can be considered as the very first step in many applications of the Semantic Web, such as sentiment classification [24], ontology generation [19, 33], query expansion [31], as well as automated knowledge delivery systems [22]. The basic techniques for automatic term extraction can be classified into four plus one (4 + 1) categories that can be summarized as following:

- **Statistical methods**: These methods use various statistical measures of words in a corpus of text to identify significant terms in a domain. Various statistical measures have been used in the literature, such as frequency of occurrence [5], mutual information (which is a measure of correlation that also refers to individual events) [10], Dice coefficient, likelihood ratio test [10], etc.
- **Allocation methods**: Distributional methods [25] are a specialized form of statistical methods that use the distribution of words in the text corpus in addition to their frequencies. An example of this method is the Term Frequency and Inverse Document Frequency (Tf-Idf) method [35].
- **Conceptual methods**: These methods are based on the assumption that the context of a word plays an important role in identifying the meaning of that word. In these methods, the internal structure along with the context of words are used to extract terms from the text. The NC-Value method [6] is an example of context-based methods.
- **Linguistic methods**: These methods identify terms based on syntactic rules. A classic example of this kind of methods constitutes part-of-speech tagging (POS tagging), stop-words and linguistic filters [6]. The basic idea behind these methods is that nouns and adjectives are probably more likely to be terms with semantic information than other grammatical types. Linguistic filters [6] can be used to identify phrases that meet certain forms, such as NP (Noun Phrase), Proposition Phrase (PP), etc.
- **Hybrid methods**: Often, a hybrid model among these aforementioned techniques is used for automatic term extraction. More to the poing, these models use a combination of two or more methods presented above [27].

2.2 Automatic Tools for the Extraction of Terms

The extraction of terms can be also utilized with use of the following three automatic tools:

(1) RAKE: Rapid Automatic Keyword Extraction (RAKE) is a simple tool for automatic term extraction [29]. The first step of this algorithm deals with document analysis; the algorithm takes as input a list of punctuation marks and stop-words and extracts them in this step. By using this step, the identification of possible keywords is limited to searching for co-occurrences of words without the need of a larger search window.

The second step consists of the score calculation regarding the keywords. Higher scores are given to keywords that are found together with other keywords and appear often together rather than words that appear frequently but alone. The third step in the RAKE algorithm is implemented when the algorithm is applied to larger texts. The algorithm identifies neighbouring keywords that appear at least three times together in the same document and in the same order. New candidate keywords are added to the list of candidate keywords by including a combination of such neighbouring keywords.
The fourth and final step of the RAKE algorithm consists of identifying the actual keywords list from the set of candidate keywords. This is done by removing the top T terms from the set of candidate keywords. Often, T is calculated as one third of the total number of candidate keywords identified by the algorithm in the previous steps.
The RAKE algorithm is available under the name rake-nltk3 in Python.

(2) TerMine: This tool uses a combination of linguistic and statistical information. Regarding linguistic methods, the POS method, the language filters as well as the stop-words list are included, while for statistical methods, the overall frequency of occurrence of the candidate word set in the corpus of text, the frequency of occurrence of the candidate word set in a larger candidate word set, and the length of the word set (in terms of word count) are included. Furthermore, this tool performs much better compared to RAKE; however, a limitation of this tool is that it can only extract word set (from 2 or more).

(3) Term Raider: TermRaider is available as a plug-in with GATE (General Architecture for Text Engineering) [19]. It requires a text corpus consisting of documents related to the field/-domain in order to extract terms. The tool makes use of linguistic and conceptual methods; more specifically, the input corpus undergoes the pre-processing step using linguistic methods such as POS tagging and stop-words removal and in following extracts the useful terms from the corpus of text based on Tf-Idf score.
TermRaider needs many documents for achieving an optimal performance. It cannot be used in case of extracting useful terms from a single domain-specific document. Moreover, a corpus of texts is necessarily required for extracting terms when using TermRaider.

2.3 Domain-Specific Term Extraction

In the field of knowledge mining, several kinds of content knowledge can be extracted from textual data (Natural Language Processing, domain-specific vocabulary and semantics among others) that can be stored in a database. Despite the strong demand from qualified researchers and experts in the area dealing with large amounts of text, there has been little practical work on the application of content extraction techniques from maritime and especially legal texts. Under this notion, the following bibliography outlines certain studies on text analysis focusing on extraction of terms from a specific domain.

Initially, Automatic Term Recognition (ATR) aims at extracting domain-specific terms from a corpus of a certain academic or technical domain. A novel idea for the automatic recognition of domain-specific terms is proposed in [23]. This specific idea is based on the statistics between a compound noun and its component single-nouns. More precisely, the paper focuses on how many nouns adjoin the noun in question to form compound nouns. Authors focus on a single-noun and a compound noun and as a next step, they use a stop-word list in order to remove. This paper indicates that technical terms along with product features are generally nouns.

Authors in [18] draw a comparative analysis of two approaches for term extraction (both linguistic and statistical) from a specific corpus regarding the pediatrics area. The linguistic based method receives a syntactically annotated corpus and extracts terms using an analysis based on the most frequent noun phrases. Each word from each phrase is annotated according to its syntactic function, its morphological characteristics and a semantic tag. The analysis of the extraction is focused only in NPs with 2 (bigrams) and 3 (trigrams) words. The second method follows a statistical approach, in which the terms are extracted through an analysis of their frequency in the corpus, and this method discards terms from a stop-words list.

Furthermore, a machine learning method to automatically classify the extracted n-grams from a corpus into terms and non-terms, is proposed in [34]. Common statistics features including TTF, X2, Tf-Idf, C-Value, RAKE for training and six machine learning algorithms, including Random Forest (RF), Linear Support Vector Machine (LinearSVC), Radial Basis Function Support Vector Machine (SVC RBF), Multinomial Naive Bayes (MNB), Linear Model Logistic Regression and Linear Model SGD Classifier, are used. The proposed method was applied to term recognition in multiple domains and languages and promising results were identified. These results indicate that this approach is capable of identifying both single and multi-word terms with reasonably good precision and recall.

Another similar work considers possible term spans within a fixed length in the sentence and predicts whether they can be conceptual terms [7]. Results show that a high recall and a comparable precision on term extraction task can be achieved. Authors contemplate the term extraction task as a progress of classifications and filtering, whereas the model can be divided into three parts: the feature preparation part, where they use the Conventional Neural Network on character level and the Long Short Term Memory Neural Network on word level, the classification part and the ranking part. On the other hand, an architecture for entity recognition in the context of the power domain is proposed in [12]. The proposed model is based on the Conditional Random Fields (CRF) as well as the bidirectional Long Short-Term Memory (LSTM). The results indicate that the CRF model outperforms the classic BLSTM model achieving an accuracy of 83%.

Similarly to our proposed work, the work in [17] conducts a literature survey with a focus on legal and Greek documents. Authors tried to describe the main aspects of the existing methodologies, i.e. Name Entity Recognition and Deep Learning techniques, recently introduced for their effective solution and the most important researches in the sub-fields of legal and Greek texts. Nevertheless,

this research has indicated the need for investigation in the area of further processing specific legal texts and is specifically aimed at solving the problems of relationship extraction, summarization and classification.

According to these previous studies, the distinctive factor of our work is the use of new concept words and therefore thematic exclusivity (i.e. nautical terms). Furthermore, our proposed application is specialized on Greek vocabulary, a subject that previous studies have not studied it on an extensive way. More specifically, we attempt to effectively evaluate and assess the classifier's output. Also, joint assessment metrics are employed in classifying and are embraced by the information retrieval, i.e. are Positive Predictive Value (Precision) and Sensitivity (Recall). However, further work has to be carried out in various directions especially on maritime terminology.

3 METHODOLOGY

3.1 Dataset

The dataset used in the experiments consists of Official Government Gazette of the Hellenic Republic (O.G.G.) corpora in the Greek language. These are legal texts on maritime topics, between 1975 and 1999. Also, the initial corpora consists of 80.000 words. These types of corpora contain common legal expressions and special characters as well, such as /, so it was necessary for data pre-processing to be performed. One of the major forms of pre-processing is to discard useless data, such as words. In Natural Language Processing, useless words are referred to as stop-words. Moreover, in linguistic morphology and information retrieval, stemming is the process of reducing inflected words to their word's stem, base or root form. We applied Python stemming library[2], which given a word, removes its suffix according to a set of rules-based algorithm. The algorithm is developed according to the grammatical rules of the Modern Greek language [11]. After the pre-processing procedure, we have calculated the number of unique words in the corpora, which is calculated to be about 4.800 words.

3.2 Feature Set

In this subsection, the feature set used is described. Text features play a crucial role in many Machine Learning tasks, such as text classification, automatic term recognition, etc.

In Natural Language Processing, features are considered important steps to be followed for a better understanding of the corpus context. After the initial text has been cleaned, it is needed to be transformed into features in order to be used by the ML model. The feature set used is based on statistical features, which are:

(1) Features based on word frequency. This set contains:
   - The simple word frequency, which is the number of times that a word appears in a corpus.
   - The word frequency in a non-topic corpus.
   - The Average Reduced Frequency (ARF), which is a variant on a frequency list that "discounts" multiple occurrences of a word, e.g. in the same document [30].
   - The Average Logarithmic Distance Frequency (ALDF), which is another type of frequency that can be displayed for the results of word lists and keyword term extraction. This modified frequency indicates whether a token is evenly distributed throughout the whole corpus or its occurrences are close to each other. The more similar ALDF is to absolute frequency, the more evenly distributed the word is. If an absolute frequency and ALDF are the same, then the word is perfectly widespread through the whole corpus. In comparison with Average Reduced Frequency (ARF), ALDF is based on distances between the words. For the calculation of these features, the Sketch Engine tool can be used [16] .

(2) Entropic measures in general are relevant for a wide variety of linguistic and computational sub-fields. In this experiment, we decided to use Shannon-entropy. This index is a measure of the degree of randomness in a set of data.

(3) Term Frequency Inverse Document Frequency (Tf-Idf) algorithm is a very popular technique, commonly used to weigh each word in the text document according to the degree of uniqueness.

3.3 Word Embeddings

Word embeddings are the dominant approach to model the words into numeric data. In the current paper, a Greek version of WordSim353 for pre-trained embeddings, is adopted. More precisely, WordSim353 contains the 300-dimensional Greek embeddings of 350K words, trained on 20M URLs with Greek language content, which were computed during 2018. More details about the number of unique sentences, unigrams, bigrams, trigrams and so forth can be found in the paper presented in [26]. In this case, the embedding layer utilized the embedding matrix produced by the embedding index dictionary and the word index as well. The size of the embedding layer, in number of nodes, is 300, that is the dimension of pre-trained embeddings.

3.4 Network Architecture

This study aims to identify if a word can be considered as a maritime term. This task can be viewed as a binary classification problem. The annotation for this problem is calculated as follows: 1 if a word is a maritime term or 0 if a word in not a maritime term. Because of the language (Greek) and the domain of the data (legal corpora), class imbalance is observed. More specifically, 25% were annotated as maritime terms and 75% as non-maritime terms.

Regarding the architecture, each term is passed to the pre-trained embedding layer producing the numerical vectors, which form the input to the hidden layers. The hidden layer outputs from the input to the output layer. In this step, an extra input to the output layer is used, i.e. the matrix $F[i, j]$, where i is the number of terms and j the number of features, containing the features (described in Section 3.2). The Network Architecture schema is illustrated in Figure 1.

The network model architecture for the experiments consists an architecture of feed-forward layers. The network is trained using the Adam optimizer [35] for optimizing parameters. To avoid overfitting, dropout is applied with a rate of 0.05, using the loss function of binary cross entropy and the regularization parameter l is set equal to $10 - 3$. 10-fold cross validation was employed in terms of testing.

[2]https://github.com/kpech21/Greek-Stemmer

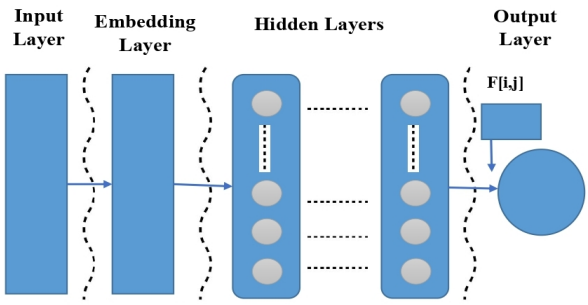

Figure 1: Network Architecture

4 EXPERIMENTAL RESULTS

In this section, the experiment results are presented and analyzed. In order to quantify and evaluate the performance of the classifier, common evaluation measures in classification are used and adopted from Information Retrieval; the Positive Predictive Value (Precision) and the Sensitivity (Recall). The first metric shows which proportion of classifications is actually correct, whereas the second metric presents the proportion of actual positives classified correctly.

Given the challenges governing the genre and the domain of the data, we have decided to apply a meta-learner to address the class imbalance (25% annotated as maritime terms and 75% as non-maritime terms) and increase robustness. The minority class (maritime terms) causes problems in the classification process; the classification algorithms give low accuracy as they tend to classify the new unseen segments in the majority class. In order to improve the performance of the model for the minority class (precision 50%, recall 20%), we used the SMOTE over-sampling approach [2] for creating new synthetic training data. SMOTE combines the feature values of minority class examples with the feature values of their nearest neighbor examples ($n = 5$), in order to produce new examples of the minority class.

Our main results are depicted in Table 1. Before using Prior SMOTE over-sampling method, the model achieves high accuracy for non-maritime terms. Concretely, Precision is equal to 74% for non-maritime terms and 50% for maritime terms, whereas Recall is equal to 50% for non-maritime terms and 20% for maritime terms. After SMOTE over-sampling, Precision is equal to 73% for non-maritime terms and 71% for maritime terms, whereas Recall is equal to 69% for non-maritime terms and 74% for maritime terms. Before using Prior SMOTE over-sampling, the model almost perfectly recognized the non-maritime terms, but it did not achieve the same good performance for the recognition of maritime terms. After the application of this method, there is a slight decrease in the detection of non-maritime terms but, at the same time, a satisfactory increase in the recognition of maritime terms, is observed.

For the purpose of comparison with other models, as proposed in [3, 4], we decided to run additional experiments based on classical algorithms, using the same feature set, utilizing the WEKA framework as back-end [9]. These experiments contain the algorithms Random Forest [1], Naive Bayes [28] and Support Vector Machines [32], with the application of the SMOTE filter. The results without applying the SMOTE filter are illustrated in Figure 2.

Table 1: Precision and Recall

Class	Precision	Recall
Model_no SMOTE		
non maritime term	74%	50%
maritime term	50%	20%
Model_with SMOTE		
non maritime term	73%	69%
maritime term	71%	74%

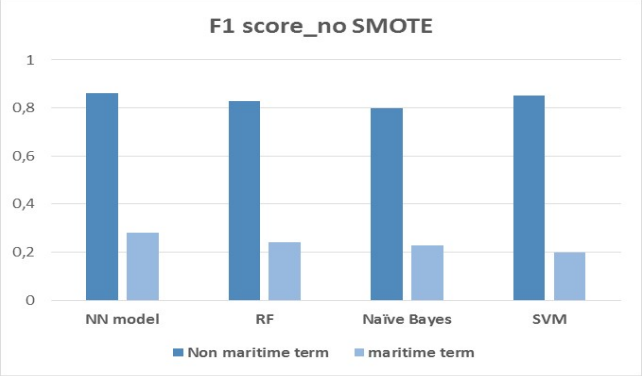

Figure 2: F1 Score Among Models_no SMOTE

The Random Forest algorithm achieved the highest performance among the classical algorithms and it had slightly worse performance than the NN model. As expected, the non-maritime terms class demonstrates the best performance when the SMOTE filter is not applied. Additionally, after applying the SMOTE filter (Figure 3), the minority class (maritime terms) has a better prediction accuracy than before. It is not observed any significant differences in the non-maritime terms for all the classifiers. Regarding the training time of the classic algorithms and the NN model, the classical algorithms have been trained faster, although it was not observed any remarkable difference compared to the NN model training time.

5 CONCLUSIONS AND FUTURE WORK

We developed a text mining technique for discovering knowledge from textual data, focusing in legal texts for extracting concepts from the broader domain of maritime heritage. As a result, our system has been challenged by the linguistic diversity of the domain-specific dataset and by adapting the traditional methodology of candidate terms formation to improve the word selection process.

A possible improvement could be to explore more complex context patterns, such as pre-training the grammatical formalism in legislative texts and to provide greater clarity in resolving complex issues that require context over a long period of time. During the application of the proposed method, some errors occurred, which are common in an original work like ours. Among these, the inconsistency of the metrics associated with word frequency, the low accuracy of the classification algorithms and the reduction in

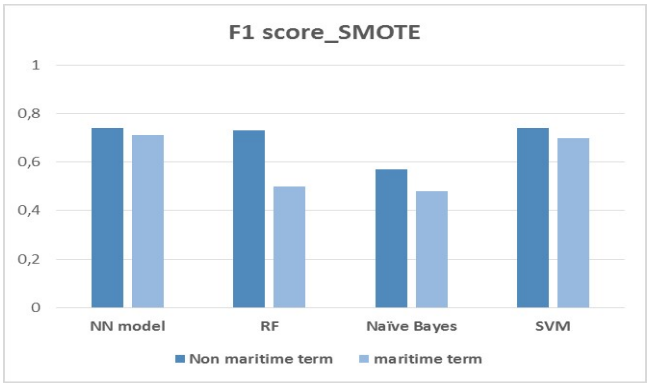

Figure 3: F1 Score Among Models_SMOTE

discoverability of non-maritime terms, can be considered. Hence, the use of appropriate pre-processing tools will help to identify the properties of the language in the corpus. It is worth mentioning that an important issue is the lack of pre-existing annotated datasets, especially for less widely used languages, such as Greek, and for specific areas, such as maritime legislation. This requires further pre-processing and significantly affects the training of the model.

ACKNOWLEDGMENTS

This research was co-financed by the European Union and Greek national funds through the "Competitiveness, Entrepreneurship and Innovation" Operational Programme 2014-2020, under the Call "Support for regional excellence"; project title: "Intelligent Research Infrastructure for Shipping, Transport and Supply Chain - ENIRISST+"; MIS code: 5047041.

REFERENCES

[1] Mariana Belgiu and Lucian Drăguț. 2016. Random Forest in Remote Sensing: A Review of Applications and Future Directions. *ISPRS Journal of Photogrammetry and Remote Sensing* 114 (2016), 24–31.

[2] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research (JAIR)* 16 (2002), 321–357.

[3] Merley da Silva Conrado, Thiago Alexandre Salgueiro Pardo, and Solange Oliveira Rezende. 2013. A Machine Learning Approach to Automatic Term Extraction using a Rich Feature Set. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics*. The Association for Computational Linguistics, 16–23.

[4] Jody Foo and Magnus Merkel. 2010. Using Machine Learning to Perform Automatic Term Recognition. In *LREC 2010 Workshop on Methods for Automatic Acquisition of Language Resources and their Evaluation Methods*. 49–54.

[5] William B. Frakes, Gregory Kulczycki, and Jason Tilley. 2015. A Comparison of Methods for Automatic Term Extraction for Domain Analysis. In *14th International Conference on Software Reuse (ICSR) (Lecture Notes in Computer Science, Vol. 8919)*. 269–281.

[6] Katerina T. Frantzi, Sophia Ananiadou, and Hideki Mima. 2000. Automatic Recognition of Multi-word Terms: the C-value/NC-value Method. *International Journal on Digital Libraries* 3, 2 (2000), 115–130.

[7] Yuze Gao and Yu Yuan. 2019. Feature-Less End-to-End Nested Term Extraction. In *8th CCF International Conference on Natural Language Processing and Chinese Computing (NLPCC) (Lecture Notes in Computer Science, Vol. 11839)*. 607–616.

[8] Panagiotis Gourgaris, Andreas Kanavos, Christos Makris, and Georgios Perrakis. 2015. Review-based Entity-ranking Refinement. In *11th International Conference on Web Information Systems and Technologies (WEBIST)*. 402–410.

[9] Mark A. Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations* 11, 1 (2009), 10–18.

[10] Maria Holmqvist, Sara Stymne, Jody Foo, and Lars Ahrenberg. 2009. Improving Alignment for SMT by Reordering and Augmenting the Training Corpus. In *4th Workshop on Statistical Machine Translation (WMT@EACL)*. Association for Computational Linguistics, 120–124.

[11] David Holton, Peter Mackridge, Irene Philippaki-Warburton, and Vassilios Spyropoulos. 2012. *Greek: A Comprehensive Grammar of the Modern Language*. Routledge.

[12] Zhixiang Ji, Xiaohui Wang, Changyu Cai, and Hongjian Sun. 2020. Power Entity Recognition based on Bidirectional Long Short-term Memory and Conditional Random Fields. *Global Energy Interconnection* 3, 2 (2020), 186–192.

[13] Kyo Kageura and Bin Umino. 1996. Methods of Automatic Term Recognition: A review. *Terminology* 3, 2 (1996), 259–289.

[14] Andreas Kanavos, Evangelos Theodoridis, and Athanasios K. Tsakalidis. 2012. Extracting Knowledge from Web Search Engine Results. In *IEEE 24th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE Computer Society, 860–867.

[15] Kush Khosla, Robbie Jones, and Nicholas Bowman. 2019. Featureless Deep Learning Methods for Automated Key-Term Extraction.

[16] Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The Sketch Engine: Ten Years on. *Lexicography* 1, 1 (2014), 7–36.

[17] Panteleimon Krasadakis, Evangelos Sakkopoulos, and Vassilios S. Verykios. 2021. A Natural Language Processing Survey on Legislative and Greek Documents. In *25th Pan-Hellenic Conference on Informatics (PCI)*. 407–412.

[18] Lucelene Lopes, Leandro Henrique M. de Oliveira, and Renata Vieira. 2010. Portuguese Term Extraction Methods: Comparing Linguistic and Statistical Approaches. In *International Conference on Computational Processing of the Portuguese Language*.

[19] Diana Maynard, Yaoyong Li, and Wim Peters. 2008. NLP Techniques for Term Extraction and Ontology Population. In *Ontology Learning and Population: Bridging the Gap between Text and Knowledge*. Frontiers in Artificial Intelligence and Applications, Vol. 167. IOS Press, 107–127.

[20] Alaa Mohasseb, Mohamed Bader-El-Den, Andreas Kanavos, and Mihaela Cocea. 2017. Web Queries Classification Based on the Syntactical Patterns of Search Types. In *19th International Conference on Speech and Computer (SPECOM) (Lecture Notes in Computer Science, Vol. 10458)*. Springer, 809–819.

[21] Despoina Mouratidis and Katia Lida Kermanidis. 2019. Ensemble and Deep Learning for Language-Independent Automatic Selection of Parallel Data. *Algorithms* 12, 1 (2019), 26.

[22] Prasenjit Mukherjee and Baisakhi Chakraborty. 2016. Automated Knowledge Provider System with Natural Language Query Processing. *IETE Technical Review* 33, 5 (2016), 525–538.

[23] Hiroshi Nakagawa and Tatsunori Mori. 2002. A Simple but Powerful Automatic Term Extraction Method. In *Coling-02: Computerm 2002: 2nd International Workshop on Computational Terminology*.

[24] Tetsuya Nasukawa and Jeonghee Yi. 2003. Sentiment Analysis: Capturing Favorability Using Natural Language Processing. In *2nd International Conference on Knowledge Capture (K-CAP)*. 70–77.

[25] Rogelio Nazar. 2016. Distributional Analysis Applied to Terminology Extraction: First Results in the Domain of Psychiatry in Spanish. *International Journal of Theoretical and Applied Issues in Specialized Communication* 22, 2 (2016), 141–170.

[26] Stamatis Outsios, Christos Karatsalos, Konstantinos Skianis, and Michalis Vazirgiannis. 2020. Evaluation of Greek Word Embeddings. In *12th Language Resources and Evaluation Conference (LREC)*. European Language Resources Association, 2543–2551.

[27] Maria Teresa Pazienza, Marco Pennacchiotti, and Fabio Massimo Zanzotto. 2005. Terminology Extraction: An Analysis of Linguistic and Statistical Approaches. In *Knowledge Mining*. Springer Berlin Heidelberg, 255–279.

[28] Irina Rish. 2001. An Empirical Study of the Naive Bayes Classifier. In *Workshop on Empirical Methods in Artificial Intelligence (IJCAI)*. 41–46.

[29] Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. Automatic Keyword Extraction from Individual Documents. *Text Mining: Applications and Theory* 1 (2010), 1–20.

[30] Petr Savický and Jaroslava Hlavácová. 2002. Measures of Word Commonness. *Journal of Quantitative Linguistics* 9, 3 (2002), 215–231.

[31] S. G. Shaila and A. Vadivel. 2015. TAG Term Weight-based N-gram Thesaurus Generation for Query Expansion in Information Retrieval Application. *Journal of Information Science* 41, 4 (2015), 467–485.

[32] Johan A. K. Suykens and Joos Vandewalle. 1999. Least Squares Support Vector Machine Classifiers. *Neural Processing Letters* 9, 3 (1999), 293–300.

[33] Paola Velardi, Paolo Fabriani, and Michele Missikoff. 2001. Using Text Processing Techniques to Automatically Enrich a Domain Ontology. In *2nd International Conference on Formal Ontology in Information Systems (FOIS)*. 270–284.

[34] Yu Yuan, Jie Gao, and Yue Zhang. 2017. Supervised Learning for Robust Term Extraction. In *International Conference on Asian Language Processing (IALP)*. 302–305.

[35] Ziqi Zhang, José Iria, Christopher Brewster, and Fabio Ciravegna. 2008. A Comparative Evaluation of Term Recognition Algorithms. In *International Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association.