Exploiting Cognitive Computing and Frame Semantic Features for Biomedical Document Clustering

Danilo Dessì[1], Diego Reforgiato Recupero[1], Gianni Fenu[1], and Sergio Consoli[2]

[1] University of Cagliari, Mathematics and Computer Science Department, Via Ospedale 72, 09124, Cagliari, Italy
{danilo_dessi, diego.reforgiato, fenu}@unica.it
[2] Philips Research, Data Science Department, High Tech Campus 34, 5656 AE Eindhoven, The Netherlands sergio.consoli@philips.com

Abstract. Nowadays, there are plenty of text documents in different domains that have unstructured content which makes them hard to analyze automatically. In particular, in the medical domain, this problem is even more stressed and is earning more and more attention. Medical reports may contain relevant information that can be employed, among many useful applications, to build predictive systems able to classify new medical cases thus supporting physicians to take more correct and reliable actions about diagnosis and cares. It is generally hard and time consuming inferring information for comparing unstructured data and evaluating similarities between various resources. In this work we show how it is possible to cluster medical reports, based on features detected by using two emerging tools, IBM Watson and Framester, from a collection of text documents. Experiments and results have proved the quality of the resulting clusterings and the key role that these services can play.

Keywords: Knowledge inference; Clustering; Biomedical documents; Cognitive Computation; Healthcare; Text-mining

1 Introduction

Data mining algorithms applied within the healthcare industry play a significant role in prediction and diagnosis of the diseases. Therefore, text and data mining concepts have widely been there employed thus saving time, money and life [2, 6, 29]. Extracting information about medication and symptoms from clinical documents has proved beneficial for the healthcare system itself [3]. Despite clinical documents are widely used for future analysis and diagnosis of the disease, it is challenging to extract useful information within clinical documents.

Clustering is the process of aggregating similar objects in groups called clusters. Clustering clinical documents provides a comprehensible summary of the collection, which gives arbitrary vision on it and allows easily detecting documents including same diseases, pathologies, etc. [7, 14, 26, 28]. The quality of

clustering is influenced also on the data features employed to perform the clustering itself. Therefore it is crucial to evaluate which features are more suitable than others to match or distinguish between various resources. The features selection process and evaluation has earned a lot of attention in past studies [4,19]. Often text documents, such as medical reports, have been managed through tf-idf (term frequency - inverse term frequency) vectors to describe their contents like in [15, 20].

With the emerging of Semantic Web resources, tools and best practices, many proposed data mining approaches combine Semantic Web technologies with data mining and knowledge discovery methods. Semantic Web resources enable to infer features which have a relevant role or meaning into an unstructured text, allowing an abstraction and a generalization of contents. The reader notices that hereafter we refer to these features as high-level features.

New systems, such as cognitive systems, allow performing analysis and rapidly obtaining useful insights out of the data. A recent cognitive system that earned a lot of attention is IBM Watson[3], which derives high-level features such as concepts, emotions, entities, keywords, relations, etc. from unstructured data, and uses advanced machine learning algorithms to derive analytics, generate predictions and hypothesis, and address question answering tasks. A recent interesting use of IBM Watson in the medical domain can be found in the joint-venture between IBM and Anthem[4], known as WellPoint, which aims to improve patient care by using the power of IBM Watson into medical platforms for identifying the most likely diagnosis and treatment options for Anthem patients[5].

WordNet [18] and FrameNet [1] are two of the most important linguistic open data resources that have been formalized several times. WordNet is a lexical database that defines synsets as groups of synonyms; each synset represents a concept, which is semantically related to other concepts through relations such as hyponomy/hypernymy, meronymy/holonymy, etc. FrameNet contains frames, which describe a situation, state or action. Each frame has semantic roles known as *frame elements* and can be evoked by Lexical Units belonging to different parts of the speech. However, its limited coverage and non-standard semantics are two major obstacles for its wide adoption in frame detection and NLP-based applications such as question answering and machine reading.

To overcome this issue, a novel frame semantic tool, Framester [11], has been recently proposed. Framester works as a graph-linked data hub between open data systems as FrameNet, BabelNet [22] and WordNet, providing a dense interlinking between existing resources and enabling a homogeneous formalization of the links through different mappings. Framester can perform semantic frames and BabelNet synsets (bnsynsets) detection which may allow a novel type of exploration of unstructured contents.

In this paper we want to propose an unsupervised and open domain approach that can perform a good quality clustering on medical reports by using a reduced

[3] https://www.ibm.com/watson/
[4] https://www.antheminc.com/
[5] https://www-03.ibm.com/press/us/en/pressrelease/35402.wss

N-dimensional model through Truncated Singular Value Decomposition, and by considering data features inferred from IBM Watson and Framester. After augmenting the medical reports with these features, we map the generated model into a reduced dimensional space, and we perform the clustering by evaluating which features produce more homogeneous clusters and give a better separation among them. We show how the use of these high-level features improves the generated clusterings and outperforms that obtained by tf-idf. This study can be employed as a base to develop tools for physicians, allowing them finding new biomedical discoveries and similar clinical cases.

More precisely, the contribution of this paper is threefold:

- We have evaluated how high-level features can be applied for matching various unstructured health text documents.
- We have compared the clustering obtained by the model built with high-level features and that obtained by the model built using baseline (tf-idf).
- We have showed how cognitive computing and semantic web tools can be employed for general analysis of unstructured and narrative data on a specific topic without previous knowledge.

This paper is organized as follows. Section 2 includes past works related to clustering medical documents and improvements of data mining techniques with the employment of semantics. Section 3 states the problem we address in this paper. Section 4 describes the medical documents we have used for clustering and the resources we have employed to infer high-level features from the medical reports. Section 5 shows how we extracted the high-level features. Section 6 describes the overall method we have proposed to cluster medical reports. Obtained results, evaluation and discussion are reported in Section 7. Section 8 concludes the paper with remarks and comments.

2 Related work

Knowledge extraction tools have appeared recently and can be applied in several applications to retrieve relevant and semantic information from texts [23, 24]. These tools are often based on statistical techniques, even though there are more recent based on open linked data and machine learning techniques. In particular, since biomedical information is being created in text form more then ever before, there is a strong need to deal with various type of electronic clinical documents through automatic processes and to find useful ways to organize them [7].

For the first time, in 2002, Yeh et al. [31] ran a text-mining competition as part of the Knowledge Discovery in Databases (KDD) Challenge Cup 2002. The task was a curation problem to evaluate medical documents from the FlyBase data set and determine whether the document should be curated based on the presence of experimental evidence of Drosophila gene products. The best performing method [27] used a set of manually constructed rules based on POS tagging, a lexicon and semantic constraints determined by examining the training documents and by focusing on figure captions, which were found to be useful

for the document clustering. In [9] the authors used a Support Vector Machine trained on the words in MEDLINE abstracts to distinguish abstracts containing information on protein-protein interactions, prior to curating this information into their BIND database. They used a bag-of-words model within their classifier and estimated that the classification system would reduce the number of abstracts that the practitioners needed to read by about two-thirds.

For classifying biomedical documents from MEDLINE, another approach was proposed in [14]. This was a semisupervised spectral method for clustering over the local-content similarities from medical documents, with two types of constraints: must-link constraints on document pairs with high (MeSH)-semantic or global-content similarities, and cannot-link constraints on those with low similarities. The authors demonstrated the performance of their approach on MEDLINE document records, outperformed linear combination methods and several well-known semisupervised clustering methods, being statistically significant [14]. Authors in [2] presented a graph-based representation for biomedical articles and used graph kernels to classify them into high-level categories. In their representation, common biomedical concepts and semantic relationships were identified with the help of ad-hoc ontologies and were used to build a graph structure with semantic features to improve clustering performance [2]. To address the growing need for efficient NLP solutions that can handle the volume and variety of clinical text, in [3] it has been developed an optimized rules-based clinical concept extractor called TRACE (Tactical Rules-based AQL Clinical Extractor), using the Annotation Query Language (AQL) and producing efficient and scalable performance on large clinical texts.

Common text mining techniques are based on the statistical analysis of a term, either word or phrase. A new concept-based mining model composed of combinations of different classic text mining approaches has been proposed in [28] to improve the text clustering quality. By exploiting the semantic structure of the sentences in documents, a better text clustering result was achieved.

Medical text processing is not a new question and the use of NLP systems on medical domain has been explored into many other works; for further details the reader is referred to the surveys in [6, 29]. Typical issues in the biomedical classification task are the lack of consistent and of well-defined syntax, and large presence of unstructured information and noise [21]. Many relevant NLP approaches have tried to overcome these issues focusing on classic entity recognition and text disambiguation techniques to create domain-specific semantic content for the analysis of medical reports [5, 21]. Today new systems, called cognitive systems, have been developed to perform NLP operations on huge amount of text. They can recognize not only the words but also their meanings and roles, allowing to contextualize the goal of the speech. One of the most famous cognitive system is IBM Watson, which is able to perform perceptions and pattern matching, make predictions and deductions, develop robotic systems, and perform more reliable NLP analysis [12]. Although it has been used in many fields, it has shown particularly good performance in healthcare applications to support physicians in their work and diagnosis. In [16] it has been used to parse

medical texts through the combination of deep linguistic learning analysis and background resources to detect and match entities and relations [16]. In [8] it has been used to build a system able to summarize the great amount of information contained into medical texts to create a new generation of Electronic Medical Records (EMR).

Following this line of research, in this paper we want to show how the use of IBM Watson and Framester can help to recognize similar reports to support the medical activity by exploiting high-level features, enabling to deal with biomedical reports without previous content knowledge, and overcoming common issues as noise and errors in unstructured clinical data.

3 Problem statement

The clustering problem we have targeted can be defined as follows: given a set of medical documents $D = \{d_1, \ldots, d_N\}$, we want to compute an assignment $\gamma : D \to \{1, \ldots, K\}$, with K being the resulting number of clusters that minimizes the objective function. The objective function is defined in terms of distance between documents. The objective is to minimize the average distance between documents and their centroids or, equivalently, to maximize the similarity between documents and their centroids. Documents are first mapped into a N-dimensional space depending on the occurrences of their high-level features. Then a Truncated Singular Value Decompositionis used to reduce the N-dimensional space and to obtain a model where each document is mapped to a numerical vector that represents its fingerprint. Fingerprints are fed to the clustering algorithms by testing both Euclidean and Cosine dissimilarity functions, and a certain number of clusters is thus generated. The metrics we have analyzed consider the effectiveness of the resulting clustering of fingerprints and how well the generated clusters include documents pertaining the same topics.

4 The input dataset

The data set used in our work was freely downloaded from the open-source iDASH repository[6]. This is characterized by a set of anonymous medical reports written in plain text. The data set is composed by 2362 English reports and each of them is characterized by specific words or a specific health domain. On the average, each report contains 400 words (the shortest document has 138 words and the longest document has 1048 words). Reports can be medical transcription samples including clinical notes, medical examination, care plans, and radiology reports of individuals. Examples of the transcriptions include admission and discharge notes, surgical transcriptions, outpatient clinical encounter, emergency visit notes, echo-cardiogram, nuclear medicine, allergies and so on.

The data set consists of a collection of non labeled medical reports. Therefore, each document may be assigned to one or more categories, but there is not

[6] https://idash-data.ucsd.edu/

explicit indication in the data set documentation of the used taxonomy or how many categories one report refers to. The category of each report lies within its file name. File names generally include the disease or/and the related body part although they do not follow precise structure or patterns. Also, it is possible that different file names have words which are synonyms and, therefore, they should be considered in the same category (e.g. there are reports concerning heart issues whose file names may have prefixes such as cardiac, heart, echo-cardiogram, cardiology etc.). The reports are various and many of them are singleton, meaning they are the only ones discussing a specific topic in the whole data set. Hence, a perfect clustering should place them as outliers.

5 Cognitive Computing and Frame Semantic services for augmenting medical reports

We have looked for services which enable to retrieve features that reduce the amount of total data and augment the overall level of information. This has been applied as the extraction of high-level features for summarization of contents of unstructured texts. We aimed to retrieve which terms and information can better distinguish two reports, through novel types of features. Table 1 shows a short comparison between the most famous cognitive services, providing an overview of their features, which can be computed on unstructured texts and employed for our purpose. IBM Watson provides more services than others. We felt that IBM Watson was the most complete and suited tool to rich our goal, since it enables to infer (i) entities which can be names of diseases, kinds of body or procedures, (ii) concepts which may not be directly written and (iii) keywords which enable to reduce the overall amount of data. In addition to IBM Watson, we have employed Framester, a novel semantic web resource that enables a mapping between various linguistic tools and an effective way to retrieve frames and bnsynsets from unstructured texts.

Table 1. Short comparison of most famous cognitive computing systems.

Ability	IBM Watson	MCS	CNLG
Sentiment	+	+	+
Taxonomy generation	+		
Concepts	+	+	
Keywords	+	+	
Entities	+		+
Intent	+	+	+
External resources	+		+

MCS = Microsoft Cognitive Service, CNLG = Cloud Natural Language Google

In the following subsections we describe the two tools that we have employed to extract semantic content from the input data set and to generate the space model.

5.1 IBW Watson

IBM Watson is a question-answering system based on advanced NLP, information retrieval, knowledge representation, automated reasoning, and machine learning. It is freely accessible from IBM Bluemix[7]. It is composed by many services such as Alchemy Language, Personality Insights, Speech to Text, and so on. Its services allow performing various tasks on different types of input; they can work on text, through vocal interaction or trade-off applications.

The service we have been using for our purposes is Alchemy Language. It can retrieve entities, concepts and keywords from texts. Entities represent people or objects that can be involved in a sentence and usually relevant roles can be assigned to them. They can have different weights that can influence goals and topics of medical report contents on distinct parts of the text. Concepts allow associating the content of a report with contexts and relevant situations that can describe a specific condition or state of patients, and might not be directly contained in the text. Keywords enable listing the content of a report generating tag clouds and allowing to focus on relevant words into the text.

Inputs of Alchemy Language service can be either structured or unstructured texts whereas output is returned in JSON format.

5.2 Framester

Framester is a novel data linked resource that works as a hub between linked open data systems as FrameNet, BabelNet and WordNet. Framester can perform semantic frames and bnsynsets detection based on the integration between NLP and semantic web techniques. It is a new frame-based ontological resource that leverages an inter-operable predicate space formalized according to frame semantics [1] and semiotics [10].

A frame is a word that is used to describe a state of something, and it can associate events of a text to specific situations. A sentence can evoke various type of aspects that can be captured by frames. A bnsynset is a set of synonymous that are aggregated to a unique meaning. Bnsynsets have an important role in semantic text analysis because allow to overcome the disambiguation problem of written texts. It is accessible through a word frame disambiguation[8] interface usable by means of command line tools.

6 The proposed method

Our work aims at performing an efficient and effective clustering of the medical reports of our input data set. The steps we have performed to achieve our goal are the following:

1. Data cleaning of the medical reports;

[7] https://console.ng.bluemix.net/
[8] https://github.com/framester/Framester/wiki/Framester-Documentation

2. Features extraction;
3. Creation of the N-dimensional space model and reduction;
4. Clustering computation.

6.1 Data cleaning of the medical reports

In this step, we cleaned all the input documents from HTML tags, and also removed semi-structured content such as tables, figures, etc. Then we matched each word against the dictionary provided by WordNet sending incorrect words and getting the correct ones. At the end, each report consisted of just English text with correct grammar and punctuation. This step was necessary as both IBM Watson Alchemy Language and Framester expect valid English text to compute their results. To compute the tf-idf values, we performed more cleaning steps. We removed numeric data, punctuation, stopwords, and transformed all text in lower case, avoiding to count more times the same word.

6.2 Extraction of features

The focus of this step was to extract the tf-idf, the high-level features (entities, keywords, concepts) through IBM Watson Alchemy Language, and frames and bnsynsets from Framester from the input data set. We developed a R script for computing the tf-idf of our corpus, and Java programs that use the APIs of IBM Watson and Framester and save the results as JSON files for quick access and reference. For the entities, two interesting filters are available to tune the extraction algorithm to the healthcare domain. These are *HealthCondition* and *Anatomy* and allowed extracting a reduced number of entities.

6.3 Creation of the N-dimensional space model and reduction

To feed data to the clustering algorithm we mapped documents as vectors representing the occurrences of the extracted features. For tf-idf, we performed a classical reduction of the matrix removing the most sparse terms and reducing the overall sparsity under the 80%. Then, we built a matrix for each type of high-level feature (frames, bnsynsets, concepts, keywords, and entities).

Let be i a report and j a feature, we generated three kinds of matrices M:

- **Binary**: meaning that if j occurs within the inferred high-level features of the document i the value $M[i, j] = 1$, otherwise $M[i, j] = 0$;
- **Weighted**: meaning that $M[i, j] = weight$, where weight has been calculated by IBM Watson and represents the influence that the high-level feature j has on report i, otherwise $M[i, j] = 0$;
- **Counted**: meaning that $M[i, j]$ contains the number of times that j occurs within the high-level features of the document i, otherwise $M[i, j] = 0$;

Since Framester does not compute weights for frames and bnsynsets, it has been possible building weighted matrices only using IBM Watson high-level features. Therefore, we have generated a total number of 14 matrices.

As the dimension of the features was very high, and the computation time was very expensive, a reduction of the number of features has been necessary in order to not incur in the well-known problem of the curse of dimensionality. The method we have adopted is the Truncated Singular Value Decomposition (Truncated SVD) as applied in previous studies like [13, 17].

Specifically, SVD decomposes the main matrix in three sub-matrices U, S and V. We have considered for clustering the product $U \cdot S$ which has the same number of rows (number of documents) and a reduced number of features (columns).

The reduction of the matrix dimension, besides decreasing the overall computational costs, allows deleting the noise that might be present into data thus deteriorating the resulting clustering. It is necessary to find a trade-off between how much to filter and how much information can be neglected for the clustering computation. It is worth nothing that different type of features can differently characterize the reduction of matrices and, consequently, the clustering results.

6.4 Clustering computation

Once we generated the reduced matrices, we used the classic hierarchical clustering for partitioning the data set since it enables performing dynamic cutting operations for evaluations of the best number of clusters. The algorithm uses the agglomerative strategy; at the beginning, each vector is assigned to a single cluster. Then the algorithm works iteratively merging at each step the pair of clusters with the minimum distance as one moves up the hierarchy.

As each hierarchical method, the resulting clustering is represented as a tree, generally known as dendrogram (see Figure 1 for an example extracted from our dataset). As indicated in Figure 1, the dendrogram can be cut at some heights depending on the number of final clusters we want to obtain. In the example we would devise three final clusters.

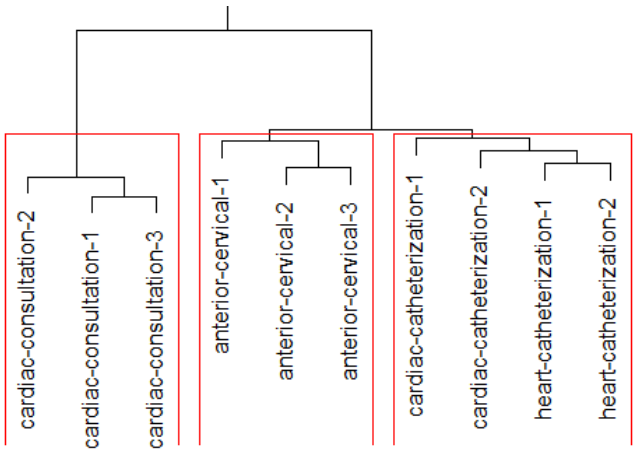

Fig. 1. Example of dendrogram on a clustering of 10 medical reports from our dataset.

In addition, we have also used the k-means algorithm for the creation of the clustering, partitioning the reports into k groups such that the sum of squares from points to the assigned cluster centers is minimized.

We have applied both clustering methods to our matrices considering both the Euclidean and Cosine similarities between the reports. We have cut the dendrogram of the hierarchical algorithm and chosen the k value of the k-means algorithm where the Silhouette width was the highest, and obtained the related resulting number of clusters.

The Silhouette width measure between two clusters is a number ranging from -1 to 1. When the value is closer to 1, it means that the clusters are well separated; when the value is closer to 0, it might be difficult to detect the decision boundary; when the value is closer to -1, it means that elements assigned to a cluster might have been assigned erroneously. In general, high average Silhouette width values between clusters indicate a good clustering. Given a cluster c, its Silhouette width $s(c)$ is computed as follows:

$$s(c) = \frac{o(c) - w(c)}{max\{o(c), w(c)\}} \tag{1}$$

where $w(c)$ is the average dissimilarity within c and the $o(c)$ is the lowest average dissimilarity of c to any other cluster.

7 Performance analysis

7.1 IBM Watson high-level features

Table 2 and Table 3 show the maximum, the average, and the standard deviation values of Silhouette width and the best number of clusters using both the Euclidean and the Cosine distances for the features extracted with IBM Watson, and performing hierarchical clustering. Similarly, Table 4 and Table 5 refer to results of k-means clustering. The reader notices that, as proven in [25, 30], Silhouette values do not reach high levels on clustering tasks related to individuals, as in our case of medical reports. This is due to the excessive requirements of domain knowledge: the question whether two patients are similar or not is always very difficult to answer, even for experienced medical experts [25, 30].

Therefore, we have used the metric based on the Silhouette width to detect the local maximums of the kernel function, whose corresponding number of clusters values provide the best quantitative separability of our data. The corresponding value for the best number of clusters has been chosen within local maximums.

For our medical reports, high-level features outperform the tf-idf in the resulting clusterings. In addition, it is worth to underline how the number of occurrences of a given entity or concept in a given report does not increase the importance of the feature itself. For an entity or a concept to be distinctive for a given report, it is enough to occur at least once. For example the entity *heart* occurs just once in several medical reports and this is enough to deduct the topic

Table 2. Results with IBM Watson, Euclidean distance and hierarchical clustering.

Feature	Type	Max	Avg	Std. Dev.	N clusters
tf-idf	-	0.026	-0.029	0.048	1497
Entities	Binary	0.308	0.257	0.026	79
Entities	Weighted	0.269	0.204	0.028	130
Entities	Counted	0.106	0.042	0.033	8
Concepts	Binary	0.281	0.248	0.036	852
Concepts	Weighted	0.274	0.218	0.026	35
Concepts	Counted	0.073	0.035	0.043	1415
Keywords	Binary	0.189	0.151	0.045	915
Keywords	Weighted	0.146	0.115	0.040	1006
Keywords	Counted	0.053	0.003	0.056	1499

Table 3. Results with IBM Watson, Cosine distance and hierarchical clustering.

Feature	Type	Max	Avg	Std. Dev.	N clusters
tf-idf	-	0.169	0.126	0.054	947
Entities	Binary	0.506	0.432	0.057	643
Entities	Weighted	0.503	0.420	0.074	37
Entities	Counted	0.245	0.200	0.034	746
Concepts	Binary	0.454	0.412	0.036	688
Concepts	Weighted	0.447	0.398	0.033	36
Concepts	Counted	0.233	0.200	0.031	874
Keywords	Binary	0.331	0.290	0.041	801
Keywords	Weighted	0.324	0.284	0.033	760
Keywords	Counted	0.219	0.177	0.045	869

of the report related to cardiac domain. This can be further observed by noticing that the binary and the weighted case for entities and concepts have higher values of Silhouette width than the counted case, indicating that the importance of entities and concepts does not really depend on their frequency within the text.

Keywords turned out to be the least distinctive and useful to perform clustering. This suggests that finding correlation between reports based only on keywords is a complex and daunting task. It is worth nothing that for Euclidean and Cosine distances the best numbers of clusters are comparable when the weighted concepts are chosen for the clustering. In fact, weighted concepts show high average values and small standard deviation values of Silhouette width.

Finally, it is possible to note how in case of Cosine similarity, entities and concepts have a similar behavior in the separation of reports into clustering especially when k-means algorithm has been adopted.

7.2 Framester high-level features

Similarly, Table 6 and Table 7 report the statistics of Silhouette width values of the high-level features extracted with Framester in case of hierarchical clustering, and Table 8 and Table 9 in case of k-means clustering. Here, the average

Table 4. Results with IBM Watson, Euclidean distance and k-means clustering.

Feature	Type	Max	Avg	Std. Dev.	N clusters
tf-idf	-	0.041	0.003	0.017	1457
Entities	Binary	0.320	0.266	0.036	842
Entities	Weighted	0.296	0.170	0.025	42
Entities	Counted	0.185	0.044	0.017	5
Concepts	Binary	0.244	0.211	0.016	45
Concepts	Weighted	0.279	0.186	0.016	48
Concepts	Counted	0.124	0.041	0.023	5
Keywords	Binary	0.150	0.123	0.020	1012
Keywords	Weighted	0.122	0.095	0.017	890
Keywords	Counted	0.232	0.021	0.026	6

Table 5. Results with IBM Watson, Cosine distance and k-means clustering.

Feature	Type	Max	Avg	Std. Dev.	N clusters
tf-idf	-	0.159	0.071	0.026	51
Entities	Binary	0.452	0.329	0.038	24
Entities	Weighted	0.514	0.350	0.051	39
Entities	Counted	0.241	0.096	0.040	33
Concepts	Binary	0.462	0.292	0.05	39
Concepts	Weighted	0.509	0.278	0.062	33
Concepts	Counted	0.201	0.093	0.038	39
Keywords	Binary	0.260	0.177	0.028	64
Keywords	Weighted	0.276	0.172	0.036	45
Keywords	Counted	0.133	0.064	0.031	74

Silhouette values are generally lower than the results obtained with the features extracted by IBM Watson. The rationale behind is that frame semantics cannot capture well the peculiarity of the health care domain where technical language is highly used. The clustering obtained by using bnsynsets as features is slightly better than that obtained with frame. Nevertheless, also these high-level features outperform the tf-idf.

Table 6. Results with Framester, Euclidean distance and hierarchical clustering.

Feature	Type	Max	Avg	Std. Dev.	N clusters
tf-idf	-	0.026	-0.029	0.048	1497
Frames	Binary	0.090	0.060	0.038	1076
Frames	Counted	0.063	0.029	0.035	1460
Bnsynsets	Binary	0.130	0.111	0.021	652
Bnsynsets	Counted	0.092	0.074	0.022	957

Table 7. Results with Framester, Cosine distance and hierarchical clustering.

Feature	Type	Max	Avg	Std. Dev.	N clusters
tf-idf	-	0.169	0.126	0.054	947
Frames	Binary	0.221	0.171	0.051	937
Frames	Counted	0.208	0.169	0.040	886
Bnsynsets	Binary	0.250	0.228	0.017	788
Bnsynsets	Counted	0.248	0.224	0.022	928

Table 8. Results with Framester, Euclidean distance and k-means clustering.

Feature	Type	Max	Avg	Std. Dev.	N clusters
tf-idf	-	0.041	0.003	0.017	1457
Frames	Binary	0.076	0.050	0.024	1236
Frames	Counted	0.111	0.023	0.023	11
Bnsynsets	Binary	0.113	0.095	0.012	1014
Bnsynsets	Counted	0.162	0.062	0.018	9

Table 9. Results with Framester, Cosine distance and k-means clustering.

Feature	Type	Max	Avg	Std. Dev.	N clusters
tf-idf	-	0.159	0.071	0.026	51
Frames	Binary	0.109	0.067	0.017	90
Frames	Counted	0.154	0.073	0.027	33
Bnsynsets	Binary	0.251	0.125	0.041	40
Bnsynsets	Counted	0.257	0.123	0.041	34

8 Conclusion

In this work, we have presented a method to cluster a set of medical reports using high-level features inferred through IBM Watson and Framester, two novel tools for cognitive computing and frame semantics. The features have been reduced using Truncated Singular Value Decomposition to solve the curse of dimensionality problem. The approach is unsupervised and does not use any prior knowledge. It is general and can be applied to other domains as well. All the developed scripts (Python, Java, R) can be requested to the authors.

Clustering on medical reports has already been covered in past research works. With the employment of semantics resources and cognitive systems, we have showed how an effective clustering of clinical reports can be obtained by the use of high-level features with respect to classical tf-idf. Combinations of high-level features might be evaluated in future together with other Semantic Web tools for improving the clustering and classification of biomedical reports.

Acknowledgments

We thank Philips Research for sponsoring the internship of Danilo Dessì in their office at High Tech Campus Eindhoven, in summer 2016.

Danilo Dessì gratefully acknowledges Sardinia Regional Government for the financial support of his PhD scholarship (P.O.R. Sardegna F.S.E. Operational Programme of the Autonomous Region of Sardinia, European Social Fund 2014-2020 - Axis III Education and training, Thematic goal 10, Priority of investment 10ii, Specific goal 10.5).

References

1. F. C. Baker, C. J. Fillmore, and J. B. Lowe. The Berkeley FrameNet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (Volume 1)*, ACL '98, pages 86–90, Stroudsburg, PA, USA, 1998. Association for Computational Linguistics.
2. S. Bleik, M. Mishra, J. Huan, and M. Song. Text categorization of biomedical data sets using graph kernels and a controlled vocabulary. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 10(5):1211–1217, 2013.
3. H. Champion, N. Pizzi, and R. Krishnamoorthy. Tactical clinical text mining for improved patient characterization. In *2014 IEEE International Congress on Big Data*, pages 683–690. IEEE, 2014.
4. Girish Chandrashekar and Ferat Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28, 2014.
5. M. Chernyshevich and V. Stankevitch. IHS-RD-BELARUS: Clinical named entities identification in French medical texts. *Physiology*, 279:291, 2015.
6. A. M. Cohen and W. R. Hersh. A survey of current work in biomedical text mining. *Briefings in bioinformatics*, 6(1):57–71, 2005.
7. S. Consoli and N. I. Stilianakis. A quartet method based on variable neighbourhood search for biomedical literature extraction and clustering. *International Transactions in Operational Research*, 24(3):537–558.
8. M. Devarakonda, D. Zhang, C.-H. Tsou, and M. Bornea. Problem-oriented patient record summary: an early report on a Watson application. In *e-Health Networking, Applications and Services (Healthcom), 2014 IEEE 16th International Conference on*, pages 281–286. IEEE, 2014.
9. I. Donaldson, J. Martin, B. de Bruijn, C. Wolting, V. Lay, B. Tuekam, S. Zhang, B. Baskin, G. D. Bader, K. Michalickova, T. Pawson, and C. W. V. Hogue. PreBIND and Textomy - Mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC Bioinformatics*, 4(1):11, 2003.
10. A. Gangemi. What's in a Schema? *Cambridge University Press, Cambridge*, pages 144–182, 2010.
11. A. Gangemi, M. Alam, L. Asprino, V. Presutti, and D. Reforgiato Recupero. Framester: A wide coverage linguistic linked data hub. In *Proceedings of Knowledge Engineering and Knowledge Management: 20th International Conference, EKAW 2016*, pages 239–254. Springer, 2016.
12. R. E. Gantenbein. Watson, come here! The role of intelligent systems in health care. In *2014 World Automation Congress (WAC)*, pages 165–168, 2014.
13. K Gayathri and A Marimuthu. Text document pre-processing with the knn for classification using the svm. In *Intelligent Systems and Control (ISCO), 2013 7th International Conference on*, pages 453–457. IEEE, 2013.
14. J. Gu, W. Feng, J. Zeng, H. Mamitsuka, and S. Zhu. Efficient semisupervised MEDLINE document clustering with MeSH-semantic and global-content constraints. *IEEE transactions on cybernetics*, 43(4):1265–1276, 2013.

15. Saeed Hassanpour and Curtis P Langlotz. Unsupervised topic modeling in a large free text radiology report repository. *Journal of digital imaging*, 29(1):59–62, 2016.
16. A. Kalyanpur and J. W. Murdock. Unsupervised entity-relation analysis in IBM Watson. In *Proceedings of the Third Annual Conference on Advances in Cognitive Systems ACS*, pages 1–12, 2015.
17. Hyunsoo Kim, Peg Howland, and Haesun Park. Dimension reduction in text classification with support vector machines. *Journal of Machine Learning Research*, 6(Jan):37–53, 2005.
18. D. Lin. Review of "WordNet: An Electronic Lexical Database" by Christiane Fellbaum. The MIT Press 1998. *Comput. Linguist.*, 25(2):292–296, 1999.
19. Huan Liu and Lei Yu. Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on knowledge and data engineering*, 17(4):491–502, 2005.
20. Filipe R Lucini, Flavio S Fogliatto, Giovani JC da Silveira, Jeruza Neyeloff, Michel J Anzanello, Ricardo de S Kuchenbecker, and Beatriz D Schaan. Text mining approach to predict hospital admissions using early medical records from the emergency department. *International Journal of Medical Informatics*, 2017.
21. M. Lushnov, T. Safin, M. Lapaev, and N. Zhukova. Medical text processing for SMDA project. In *EMSA-RMed@ESWC*, 2016.
22. Roberto Navigli and Simone Paolo Ponzetto. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, 2012.
23. V. Presutti, S. Consoli, A. G. Nuzzolese, D. Reforgiato Recupero, A. Gangemi, I. Bannour, and H. Zargayouna. Uncovering the semantics of wikipedia pagelinks. In *Lecture Notes in Computer Science*, volume 8876, pages 413–428, 2014.
24. V. Presutti, A. G. Nuzzolese, S. Consoli, A. Gangemi, and D. Reforgiato Recupero. From hyperlinks to semantic web properties using open knowledge extraction. *Semantic Web*, 7(4):351–378, 2016.
25. B. Qian, X. Wang, N. Cao, H. Li, and Y.-G. Jiang. A relative similarity based method for interactive patient risk prediction. *Data Mining and Knowledge Discovery*, 29(4):1070–1093, 2015.
26. D. Reforgiato Recupero. A new unsupervised method for document clustering by using wordnet lexical and conceptual relations. *Information Retrieval*, 10(6):563–579, 2007.
27. Y. Regev, M. Finkelstein-Landau, and R. Feldman. Rule-based extraction of experimental evidence in the biomedical domain: The KDD Cup 2002 (task 1). *ACM SIGKDD Explorations Newsletter*, 4(2):90–92, 2002.
28. S. Shehata, F. Karray, and M. Kamel. An efficient concept-based mining model for enhancing text clustering. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1360–1371, 2010.
29. R. Toor and I. Chana. Application of IT in Healthcare: A systematic review. *ACM SIGBioinformatics Rec.*, 6(2):1–8, 2016.
30. F. Wang, J. Sun, and S. Ebadollahi. Composite distance metric integration by leveraging multiple experts' inputs and its application in patient similarity assessment. *Statistical Analysis and Data Mining*, 5(1):54–69, 2012.
31. A. S. Yeh, L. Hirschman, and A. A. Morgan. Evaluation of text data mining for database curation: Lessons learned from the KDD Challenge Cup. *Bioinformatics*, 19 Suppl. 1:331–339, 2003.