Project Title: Advanced Multi-Label Learning Techniques (AMULET)

Project ID: HFRI-FM17-514

Principal Investigator: Grigorios Tsoumakas

Host Institution: Aristotle University of Thessaloniki

Project Website: `https://amulet.csd.auth.gr`

Deliverable D2.3 – Technical Report on Dealing with Concept Drift

Team AMULET
Grigorios Tsoumakas
Stamatis Karlos
Nikolaos Mylonas

External Collaborators
Eirini Papagiannopoulou

1 Introduction

This report documents our work on examining the drifting of MeSH (Medical Subject Headings) descriptors throughout the different years of the vocabulary and specifically from 2013 to 2019. We performed a twofold analysis on MeSH descriptors, one based on the performance of a state-of-the-art model on each descriptor, and one based on the usage change of each descriptor. Doing so we aim to identify the descriptors whose semantic meaning changes, either due to changes in the ontology or due to the passing of time. The data used for our experiments come from the BioASQ challenge datasets, which are publicly available for use and contain biomedical articles indexed with the MeSH vocabulary. The two methods for drift detection along with the process of creating the final datasets used during our experiments can be found in the following sections.

2 Performance Based Semantic Shift Detection

In order to examine whether or not the phenomenon of concept drift appears in MeSH, we decided to evaluate the performance of a state-of-the-art machine learning classifier for text classification on data annotated using different versions of the thesaurus. The classifier we decided to use for this procedure is the well known language model BERT[1], which has produced very competitive results on text classification tasks in the past. For our experiments, we decided to train the selected classifier with data corresponding to the first available MeSH year in the BioASQ challenges, namely 2013. The trained classifier was then used to predict the data for the rest of the years. This way we can evaluate the performance of the model each year and catch drifts corresponding to changes in meaning, during the year where the drift occurred. In order to determine whether a descriptor has drift in meaning or not we plot the differences in F1-score between consecutive years for each descriptor and examine these plots in order to find the descriptors that act as outliers. These descriptors are the ones we consider as having drift in meaning for that year pair.

2.1 Dataset creation process

The main focus of our experiments was to see if MeSH descriptors drift throughout the years. As such, the dataset for each year had to contain articles introduced during that specific year. In order to do that, we mined each BioASQ dataset, looking only for the articles that specifically mention the year we are interested in. Since each BioASQ dataset is introduced at the start of the corresponding year of the challenge, the articles that actually refer to that year are very limited (in some cases less than 1000, when the whole dataset contains around 10 million articles), we use the BioASQ dataset version that corresponds to the year right after the one we want to create the dataset for. This way the number of articles that actually refer to that year is much higher making it possible to create datasets with sufficient information. With the above in mind we

used the BioASQ datasets corresponding to years 2014-2020 in order to create the ones we used for our experiments that refer to years 2013-2019.

After we obtained the data for each year we had to narrow down the descriptors we would focus our experiments on for computational reasons, as each dataset had around 10,000 descriptors. To alleviate this issue we decided to focus on a subset of those descriptors and specifically the most frequent ones in each year's dataset. We first removed the top 10 most frequent descriptors for each year, as these descriptors have very general meaning that holds very little information and are not particularly useful. Examples of such descriptors are *Humans*, *Male* and *Female* that appear whenever an article refers to humans, and *Mice* that is very commonly used in various experiments. We then decided to keep the 300 most frequent descriptors for each year, obtaining 7 descriptors sets. In order to study whether a descriptor's meaning drifts or not, that descriptor has to be part of all 7 datasets. Therefore, we took the intersection of the 7 descriptor sets as our examined descriptors, leaving us with 198 descriptors. The size of each year's dataset can be found in Table 1, where we show the number of articles available each year along with how many are left after the selection of the 300 and 198 descriptors.

	Number of Articles		
Year	All	top 300	top 198
2013	541,024	317,446	304,116
2014	368,527	241,628	231,231
2015	243,035	147,611	141,980
2016	129,998	119,896	113,155
2017	224,424	206,901	196,326
2018	224,608	205,897	191,580
2019	237,079	215,195	197,180

Table 1: Dataset Sizes

2.2 Results

In Figure 1 we show the plots for the F1 difference between each year pair. For each plot the x axis corresponds to the 198 descriptors with each one given an identifier from 0 to 197, while the y axis corresponds to the difference in F1-score for that descriptor between these 2 years.

We can see from these plots that most of the descriptors are grouped together, meaning they have similar differences in F1-score. The descriptors outside of that group in each plot are the ones whose behavior differs from the rest and as such are considered outliers, thus are the ones we can consider as having drift in meaning.

In Table 2 we show the mean difference value for F1-score between each one of these year pairs. The mean values are signed numbers, with a positive sign denoting that the mean F1 has increased between these years, while a negative

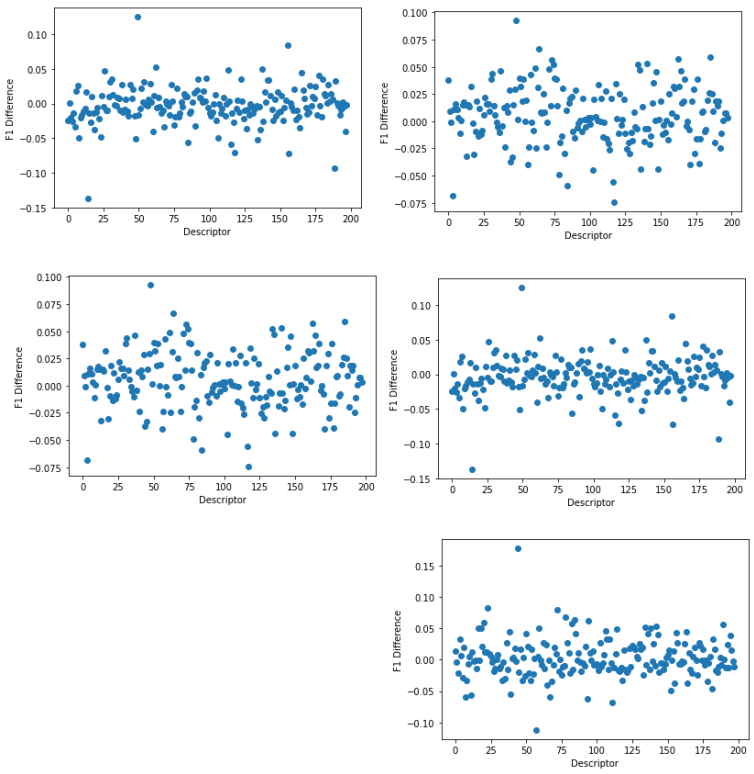

Figure 1: F1-Difference plots for each year pair. Top Left 2014-2015, Top Right 2015-2016, Middle Right 2016-2017, Middle Left 2017-2018, Bottom Right 2018-2019.

one shows a decrease.

Year-Pair	Mean Value
2014-2015	-0.0005
2015-2016	-0.0009
2016-2017	-0.0028
2017-2018	0.0055
2018-2019	0.0031

Table 2: Mean Difference Values

Based on the plots as well as their difference to mean value for each year pair we show in Table 3 the descriptors with the biggest drift in meaning for each available year pair, along with their computed difference value.

Year	F1-diff	Descriptor
2014-2015		
	0.0718	Chromatography, High Pressure Liquid
	-0.0791	Species Specificity
	-0.0753	Rats
	-0.0730	Recombinant Proteins
2015-2016		
	0.111	Base Sequence
	-0.121	RNA, Ribosomal, 16S
	-0.099	DNA, Bacterial
2016-2017		
	0.125	DNA, Bacterial
	-0.137	Base Sequence
2017-2018		
	0.092	DNA
	-0.074	Models, Theoretical
	-0.068	Aging
2018-2019		
	0.177	Computational Biology
	-0.112	Disease-Free Survival

Table 3: Most drifting descriptors per year pair

From Table 3 we can see that 2 descriptors appear in different years specifically *DNA, Bacterial* and *Base Sequence*. The former shows a decrease in 2015-2016 and an increase in 2016-2017. We could not find any changes in MeSH for the first descriptor during either of those years that justify this change but the number of appearances for that descriptor is really small in the 2016 dataset (619), while that number increased to about 3 times as much in 2017 (1706), which may be the reason for that change in performance. In case of the latter descriptor, namely *Base Sequence*, this descriptor showed an increase during 2015-2016 and then a decrease in 2016-2017. This decrease can be attributed to

the fact that during 2017 the indexing policy for descriptor *Molecular Sequence Data*, which is related to our examined, changed to only include general articles about sequence data. As a result the changes during 2017 indirectly affected the descriptor *Base Sequence* which can explain that decrease in performance. Furthermore the descriptor *Rats*, which shows a decrease during 2014-2015, had its indexing policy changed during 2015, which may have negatively affected the performance of the model on said descriptor. Specifically the descriptor *Rats*, which has a very general meaning, should no longer be used on articles where one of its children terms is selected by the indexer.

2.3 Analysis for percentage difference

Since absolute difference does not always gives us the true picture of the difference between F1 scores since higher scores will exhibit higher differences, we decided to do the exact same analysis for our descriptors this time using the percentage difference between each year. With this in mind, in Figure 2 we show these quantitative differences and in Table 4 we show the percentage mean difference for each year pair.

Year-Pair	Mean Value
2014-2015	-0.39%
2015-2016	0.53%
2016-2017	-0.93%
2017-2018	1.92%
2018-2019	1.34%

Table 4: Percentage mean difference values.

Finally in Table 5 we show the most drifting descriptors for each year pair based on the aforementioned analysis.

We can see from the above table that the results are pretty similar to those from Table 3, with the biggest difference being for years 2017-2018 where the only descriptor that remains in both Tables is *Models, Theoretical*. What stands out the most though is the descriptor *Computational Biology*, which shows a small decrease in score (17.07%) for years 2015-2016 but a substantial increase (129.12%) in years 2018-2019. After studying the MeSH changes for these specific years we could not find anything related to this descriptor. Furthermore the frequency of the same descriptor does not show any significant increase or decrease in these year pairs.

3 Usage-Based Semantic Shift Detection

3.1 Methodology

In this part, we follow the approach of [2] to detect words that differ in their usage between the corpora described in detail in Section 2.1. Specifically, we

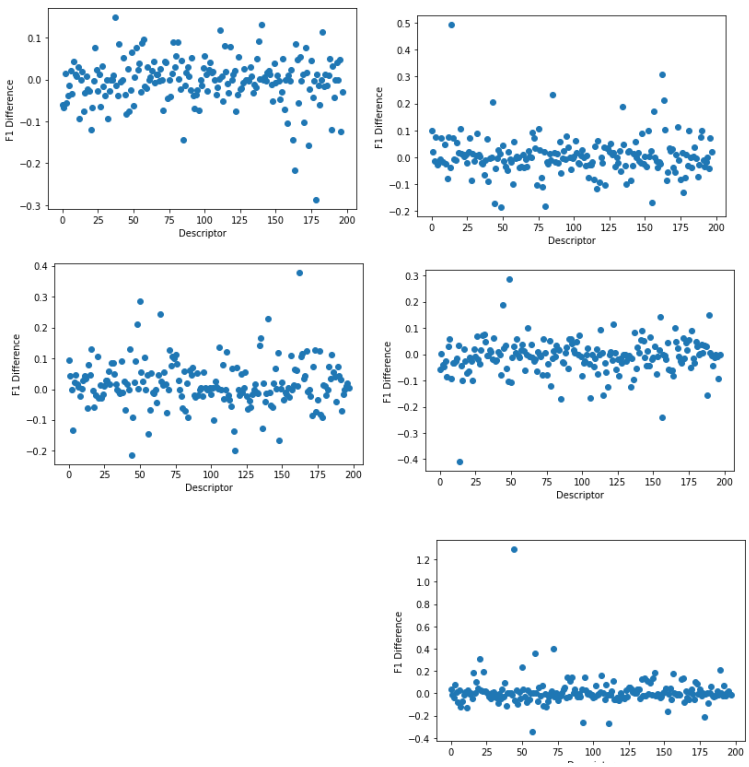

Figure 2: Percentage F1 difference plots for each year pair. Top Left 2014-2015, Top Right 2015-2016, Middle Right 2016-2017, Middle Left 2017-2018, Bottom Right 2018-2019.

Year	F1-diff	Descriptor
2014-2015		
	11.74%	Mice, Nude
	13.06%	Polymerase Chain Reaction
	14.84%	Chromatography, High Pressure Liquid
	-28.69%	Species Specificity
	-21.68%	Recombinant Proteins
	-15.74%	Sex Factors
2015-2016		
	49.24%	Base Sequence
	-18.46%	DNA, Bacterial
	-18.08%	HeLa Cells
	-17.07%	Computational Biology
	-16.83%	RNA, Ribosomal, 16S
2016-2017		
	28.55%	DNA, Bacterial
	-40.86%	Base Sequence
2017-2018		
	37.76%	Real-Time Polymerase Chain Reaction
	-21.32%	Computational Biology
	-19.81%	Models, Theoretical
	-16.45%	Proportional Hazards Models
2018-2019		
	129.12%	Computational Biology
	-34.4%	Disease-Free Survival

Table 5: Most drifting descriptors per year pair based on quantitatitve difference

define the task as follows: given two different corpora with overlapping vocabularies, identify words that their use is different in the two corpora. We return a ranked list of words, from the one that is most likely to have usage-change, to the least likely one. We represent each word in a corpus as the set of its top k nearest neighbors. We then compute the score for word usage change across corpora by considering the size of the intersection of the two sets. Our code is available here[1].

3.2 Experimental Study

First, we find the complete set of MeSH descriptors used in the corpora for all years and convert them to lowercase, concatenating the words (tokens) of descriptors that constitute multiple terms (by replacing commas and spaces with a hashtag character, e.g., *Cells, Cultured* $\rightarrow$ *cells#cultured*). Then, we pre-process each BioASQ dataset by converting it to lowercase and removing the following punctuation marks:

$$',"!,.:;?()[]$$

Next, we concatenate the texts of each year in one string using a change line character for each distinct text. Finally, we train a GloVe model on each collection (one GloVe model for the texts of the same year). We use 50-dim GloVe vectors with 10 words context window. We perform frequency-based filtering of the vocabulary, removing words with less than 5 occurrences in each corpus. We do not perform any other form of filtering.

		Frequency in corpus	
Descriptor	Common neighbors	2014	2015
Bacterial Proteins	0	31	16
Cell Movement	0	55	38
Computational Biology	0	38	20
Protein Conformation	0	32	22
Kaplan-Meier Estimate	1	21	18
Protein Binding	1	209	106
Antineoplastic Agents	3	12	13
Oxidation-Reduction	4	34	17
Cells, Cultured	7	272	169
Plant Extracts	7	165	87

Table 6: Top 10 descriptors' semantic shifts for 2014-2015.

Regarding the approach used [2], we consider neighbors and descriptors that appear in both corpora and have a raw frequency greater than 10 in the corpus under consideration. We identify the 50 nearest neighbors ($k = 50$) for each

[1] `https://drive.google.com/drive/folders/1JEnQYWutCbFhythp_iWSLkhHx5kuNyAl?usp=sharing`

descriptor to perform the intersection. We use a lower number of nearest neighbors and word frequency thresholds compared to the original work of [2] as we work on smaller corpora.

		Frequency in corpus	
Descriptor	Common neighbors	2015	2016
Antineoplastic Agents	0	13	11
Protein Conformation	0	22	14
Bacterial Proteins	1	16	11
Computational Biology	1	20	13
Cell Movement	2	38	30
Diabetes Mellitus, Type 2	3	11	13
HIV Infections	3	46	56
Molecular Structure	3	95	63
Environmental Monitoring	5	37	32
Oxidation-Reduction	5	17	26

Table 7: Top 10 descriptors' semantic shifts for 2015-2016.

Tables 6 - 10 show the top 10 descriptors' semantic shifts for each dataset pair (2014-2015, 2015-2016, 2016-2017, 2017-2018, 2018-2019)[2]. The first column is the descriptor's name, the second column shows the number of common neighbors for each descriptor in the corresponding pair of datasets (e.g., 2014-2015, etc.), and the third column shows the frequency of each descriptor in each dataset. For example, Table 6 shows that the descriptor *Bacterial Proteins* has 0 neighbors in common between the set of neighbors based on the GloVe vectors of the 2014 dataset and the set of neighbors based on the 2015 dataset, whereas the descriptor *Cells, Cultured* has 7 neighbors in common between the corresponding sets of neighbors.

		Frequency in corpus	
Descriptor	Common neighbors	2016	2017
Antineoplastic Agents	0	11	20
Bacterial Proteins	0	11	23
Cell Movement	1	30	47
Protein Conformation	1	14	33
Computational Biology	2	13	29
Oxidation-Reduction	2	26	30
Molecular Structure	3	63	99
HIV Infections	4	56	86
Protein Binding	6	91	173
Diabetes Mellitus, Type 2	7	13	31

Table 8: Top 10 descriptors' semantic shifts for 2016-2017.

[2]Note: We could show in a table the different neighbors' sets for one/two descriptors (e.g., the top 10 neighbors per year) highlighting the different contexts. However, we need a domain expert to provide a deeper interpretation regarding the actual descriptors' semantics shifts based on the corresponding neighbors' sets over time.

Descriptor	Common neighbors	Frequency in corpus 2017	2018
Antineoplastic Agents	0	20	19
Colorectal Neoplasms	0	21	13
Bacterial Proteins	1	23	17
Diabetes Mellitus, Type 2	1	31	38
Protein Conformation	1	33	29
Cell Movement	2	47	21
Computational Biology	2	29	38
Species Specificity	4	22	11
Protein Binding	5	173	166
Cells, Cultured	7	199	164

Table 9: Top 10 descriptors' semantic shifts for 2017-2018.

Descriptor	Common neighbors	Frequency in corpus 2018	2019
Antineoplastic Agents	0	19	24
Bacterial Proteins	0	17	19
Colorectal Neoplasms	0	13	19
Cell Movement	1	21	20
Protein Conformation	1	29	30
Computational Biology	4	38	26
Oxidation-Reduction	4	56	78
Species Specificity	4	11	13
Computer Simulation	7	58	46
Protein Binding	8	166	196

Table 10: Top 10 descriptors' semantic shifts for 2018-2019.

4 Comparison with Concept Drift Results

This section focuses on the semantic shift of the most drifting descriptors per year pair presented in Table 3. Specifically, Table 11 provides the results according to the approach of [2] for the most drifting descriptors presented above. The first column shows the year pair, the second column gives the descriptor's name, and the third column briefly describes the corresponding results.

We can see in Table 11 that for the majority of the descriptors presented, their number of appearances in the corpus is very small or even zero and as such this method cannot export concise results for them. For the rest of them we could not find any semantic shift using the aforementioned method. It is worth noting that this is a preliminary research and as these results should not be considered conclusive, more research is needed in order to obtain robust results about the drifting of MeSH descriptors.

Time period	Descriptor	Description
2014-2015	Chromatography, High Pressure Liquid	Not appear in the corpus' vocabulary due to the preprocessing we follow.
	Species Specificity	Frequency less than 10 in 2015 corpus.
	Rats	It seems that there is no semantic shift (38 common neighbors out of 50).
	Recombinant Proteins	It seems that there is semantic shift (13 common neighbors out of 50).
2015-2016	Base Sequence	Not appear in the corpus' vocabulary due to the preprocessing we follow.
	RNA, Ribosomal, 16S	Not appear in the corpus' vocabulary due to the preprocessing we follow.
	DNA, Bacterial	Not appear in the corpus' vocabulary due to the preprocessing we follow.
2016-2017	DNA, Bacterial	Not appear in the corpus' vocabulary due to the preprocessing we follow.
	Base Sequence	Not appear in the corpus' vocabulary due to the preprocessing we follow.
2017-2018	DNA	It seems that there is no semantic shift (34 common neighbors out of 50).
	Models, Theoretical	Not appear in the corpus' vocabulary due to the preprocessing we follow.
	Aging	It seems that there is no semantic shift (34 common neighbors out of 50).
2018-2019	Computational Biology	It seems that there is semantic shift (4 common neighbors out of 50).
	Disease-Free Survival	It seems that there is no semantic shift (34 common neighbors out of 50).

Table 11: Examining semantic shift for the most drifting descriptors presented in Table 3.

References

[1] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2019). arXiv:1810.04805.

[2] H. Gonen, G. Jawahar, D. Seddah, Y. Goldberg, Simple, interpretable and stable method for detecting words with usage change across corpora, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 538–555. doi:10.18653/v1/2020.acl-main.51.
URL `https://aclanthology.org/2020.acl-main.51`