Distributional Semantics for Medical Information Extraction

Quiroz, L.; Mennes, L.; Dehghani, M.; Kanoulas, E.

Publication date
2016

Document Version
Final published version

Published in
CLEF 2016 : CLEF2016 Working Notes

License
CC0

Link to publication

Citation for published version (APA):
Quiroz, L., Mennes, L., Dehghani, M., & Kanoulas, E. (2016). Distributional Semantics for Medical Information Extraction. In K. Balog, L. Cappellato, N. Ferro, & C. Macdonald (Eds.), *CLEF 2016 : CLEF2016 Working Notes: Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum : Évora, Portugal, 5-8 September, 2016* (pp. 109-122). (CEUR Workshop Proceedings; Vol. 1609). CEUR-WS. http://ceur-ws.org/Vol-1609/16090109.pdf

Distributional Semantics for Medical Information Extraction

Lautaro Quiroz[1]
lautaro.quiroz@student.uva.nl

Lydia Mennes[2]
lydia@ctcue.com

Mostafa Dehghani[1]
dehghani@uva.nl

Evangelos Kanoulas[1]
e.kanoulas@uva.nl

[1] University of Amsterdam, Amsterdam, The Netherlands
[2] CTcue, Amsterdam, The Netherlands

Abstract. This report describes two methods implemented for the CLEF eHealth 2016 Task 1 challenge. They consist of: a) a feed forward neural network; and b) a random forest for classification and a feed forward neural net, applied to automatically fill in medical handover forms using synthetic medical records as inputs. Both approaches are interesting because they rely on word embeddings, are domain independent, and are feature engineering free. We discuss the complexity of the task, and the impact in our models, having too many output classes and a limited amount of training data. The performance of the methods are based on traditional classification metrics (e.g. precision, recall, and f1-score) on the macro and micro averaged level, and focus on two sets of labels: a) the *"NA"* tag, which recognize data that is irrelevant and therefore should be excluded from the form; and b) all other tags, which account for the different fields of the form. The neural network achieved an F1-score of 0.8 (for the *"NA"* tag) and a macro-averaged F1-score of 0.308 and a micro-averaged result of 0.514 (for the remaining categories), while the ensemble pipeline got 0.813 (for the *"NA"* tag) and 0.345 and 0.503 for the macro- and micro-averaged rates on the rest of the labels.

1 Introduction

This year's CLEF eHealth challenge consists of 3 tasks[1], that cover different ongoing research topics, briefly summarized as: *1 - Information extraction*; *2 - Multilingual Information extraction*; *3 - Information retrieval.* Due to several factors, such as the overloading amount of information and the lack of standardization procedures when documenting cases, the information flow in the clinical field results hindered. In consequence, not only non-medical personnel but clinicians have problems in processing this information. Ultimately, these tasks would help in the way in which medical records are handled, processed, and shared, leading to a better understanding overall.

This report describes two statistical approaches to solve the first of these tasks: *Task 1: Handover information extraction.*[2]. In this first assignment, we are presented with plain text records which are the result of automated speech recognition translations from nurses' shifts verbal information exchange, and we are asked to identify relevant chunks in order to complete a clinical handover form in a fully-automated fashion.

The outline of this report is set as follows: In section 2, a description and analysis of the datasets and the methods is presented; in section 3, the experiments results are shown and explained; section 4 includes conclusions drawn from this work, as well as encountered issues and future work; finally, there is an appendix section that includes additional information referenced in the report.

2 Methodology

2.1 Datasets

Three datasets were released for the purposes of this challenge[3][4]. Though it was not compulsory, there were meant to be used as independent: a) training; b) validation; and c) testing sets. This was the case for all experiments described in this report, and that is how they are going to be referenced from now on.

Table 1 presents an overview of the datasets. These numbers account for tokens found in the data, with punctuation removal[3] as the only pre-processing step applied. As it is shown in the table, the datasets are roughly the same in terms of size, namely: number of records included, number of tokens, and number of word types. But nearly half of the word types present in the validation and testing set are not seen in the training group (50.04% and 56.55%, respectively, when considering stopwords, and 56.62% and 62.65% excluding them). This is, definitely, an obstacle to overcome; hopefully, the vector representations are going to capture enough semantic meaning to deal with it.

Note that, with no prior handling, constructions like: *'forty-eight', 'self-caring', 'self-inflammatory'*; numbers, such as: *'81', '220'*; symbols and punctuation: *'@',',','.'*; misspellings and wrongly spaced words: *'bed2', '1pm', 'gout.', 'litres/nasal', 'urine.and', 'gastroscopy/colonoscopy'* are treated as single tokens. This criterion follows the way in which the dataset was originally tokenized and listed with the given features.

Already at this stage it can be pointed out, one of the most important limitation to this report's approaches, and to other techniques of the same nature: there is not enough training data in order to effectively train the neural networks. Considering the number of parameters to be learnt, these models might result too complex to train.

[3] The following symbols were left, as there is a vector representation for them: *'_', '∧', '@','=','>','*','+','&','$','%','#'*

Table 1: Datasets overview

Dataset	# docs	# tokens	# word types	Token overlap w/stopwords	Token overlap w/o stopwords
Training	101	7451	1347	-	-
Validation	100	6798	1291	645 (49.96%)	560 (43.38%)
Testing	100	5741	1213	527 (43.45%)	453 (37.35%)

In this task, we are going to learn patterns from the data so as to fill in a fixed handover form. The description of this form can be found in the dataset paper, for the purposes of this explanation, it is relevant to know that it consists of 36 tags/labels; one of these is the *'NA'* label, which accounts for information that shouldn't be included in the form. Naturally, this tag covers the most number of tokens, and as we are dealing with a multiclass classification task, this difference in label-group sizes will play a significant role at prediction time, specifically when computing the resulting averaged metrics.

Looking at the training and validation sets, we can find out that there is a mismatch between the tags that are included in one and the other. Namely, the training dataset includes 36 labels, 3 of which are not found in the validation set (tags: *'Appointment/ Procedure_ClinicianGivenNames/ Initials'*, *'Appointment/ Procedure_Ward'*, *'Appointment/ Procedure_City'*); and the validation set includes 36 tags as well, 3 of which are not seen in the training set (tags: *'PatientIntroduction_Title'*, *'Appointment/ Procedure_ClinicianTitle'*, *'Appointment/ Procedure_Hospital'*). Finally, the testing set is in agreement with the training label data. This difference will have an impact on the validation scores, but not on the testing ones.

Conceptually, the statistical models to be trained will try to fit the conditional probability of a label given certain word tokens. It is worth analysing the data at the tag level so to get an idea of how complex the task is. Table 5 and Table 6, included in the Appendix, show the empirical distribution shaped from the training data, and a word type count overlap breakdown. The overlap summary was done considering the validation and the training set (at the moment of writing this report, the testing set labels were not released), and it includes stopwords in the counts. As it can be seen in the empirical distribution, there is one category which concentrates the majority of the tokens (the *"NA" tag*, leaving a very low probability mass for the rest of the labels. Clearly, some categories are easier than others to predict; but, in the end, this is going to be dependant on: a) the number of samples; and b) the token overlap of the category. These two factors will affect the quality of the trained word embeddings.

In the next subsections, the details of the two models are provided.

2.2 Method 1: a simple neural network approach

The first approach consists of a one hidden layer context window feed forward neural network. Following the idea of constructing a pipeline that is domain independent, there are no features derived from medical data enrichment, and the neural network makes use of semantic features only, i.e. word embeddings.

From a Natural Language Processing point of view, we know that languages present a high level of ambiguity, and if we, furthermore, take into account the overlap schema presented in the previous section, it seems like a good idea to include a context window on the token to be predicted.

As mentioned before, even this simple architecture might have too many parameters to be trained with the amount of data we have. As a way to help this situation, the word embeddings are initialized using the pretrained Googlenews Word2Vec representations[4][5].

Figure 1 shows a graphical description of the implemented neural network[9].

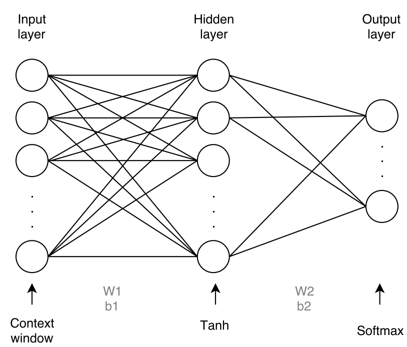

Fig. 1: Neural network architecture

The network has two weight matrix and two bias parameters. The matrix shown as *W1* corresponds to the word embeddings, initialised uniformly and intersected with the pretrained embeddings, while the matrix *W2* and the bias vectors *b1* and *b2* are uniformly initialised. There is a non-linear behaviour in the net, introduced by the hyperbolic tangent function on the hidden layer, and the output layer is a classification step achieved by using a softmax activation function. The output layer has a dimension of 39, because of the 36 tags in the training/test set and the 3 added tags found in the validation set, as explained in the previous section.

The neural network is trained with Stochastic gradient descent, and back-propagation[6], using Adagrad[7] as gradient update optimizer. Concerning the overfitting behaviour of neural nets, an L2-regularization is applied to the weight matrices *W1* and *W2*.

[4] `https://code.google.com/archive/p/word2vec/`

2.3 Method 2: an ensemble approach

In this second method, the pipeline is based on a layered prediction. First, a random forest[8] is set up with the purpose of predicting a subset of the tags, and then a neural network, with the same architecture as described in the previous subsection, is implemented to further discriminate between the remaining labels.

Again, both, the random forest as well as the neural network only make use of semantic features. It could be the case that the two models use different or extended types of features in order to increase their performance. As an advantage, these two methods could possibly learn different patterns from the data, helping the final prediction quality. But, at the same time, the errors that are made in the first step, are further passed to the second model; because false-positive predictions by the random forest are taken out from the sample set to be predicted later by the neural net, and false-negatives are going to be included by the neural net in one of the remaining categories as false-positives values.

The random forest uses decision trees for classification, with 100 trees and a maximum depth of 5. The neural network is trained as described in method 1.

2.4 Pretrained word embeddings

The neural networks described above utilize pretrained word embeddings as features. Given the small amount of data for training, and considering the novelty, with respect to word types, that are included in each dataset, if a word has not been seen in the training data and has no representation in the Googlenews pretrained embeddings, then it will be assigned a randomly generated vector. In these cases, there is no semantic information to use so as to assign a class.

Table 2 shows the statistics of the previously described cases for word tokens and word types in the validation and testing set.

Table 2: Random word representations

Dataset	**Tokens**		**Word types**		**Most affected tags (top 3)**
	#	%	#	%	**(tag: %)**
Validation	231	0.03	160	0.114	PatientIntroduction_Lastname: 0.156 PatientIntroduction_GivenNames/Initials: 0.1 MyShift_OtherObservation: 0.091
Testing	782	0.12	602	0.475	*Tags not available*

3 Results and analysis

The performance of the methods measure the precision, recall, and F1-score using the *conlleval evaluation script*, as implemented in the CoNLL 2000 Shared Task on Chunking[5].

[5] http://www.cnts.ua.ac.be/conll2000/chunking/

In this section, the results for the two methods are presented and compared.

At the moment of submission, an error when writing the output file produced an alteration in the order of the predicted tags, completely mixing the results. The associated scores are not included in this report.

3.1 Method 1

Table 3 shows a summary of the results obtained when using the neural network with a context window size of 7, after being trained for 50 epochs. A detailed, per tag, analysis can be found in the Appendix section, in Table 7.

Table 3: Method 1 results

Dataset	**Macro average**			**Micro average**			**NA**		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Training	0.741	0.591	0.624	0.908	0.862	0.884	0.92	0.979	0.948
Validation	0.468	0.344	0.355	0.636	0.495	0.557	0.696	0.92	0.793
Testing	0.411	0.307	0.308	0.563	0.472	0.514	0.723	0.894	0.8

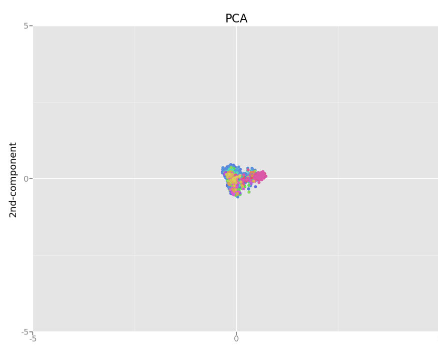

Fig. 2: Word embeddings PCA before training

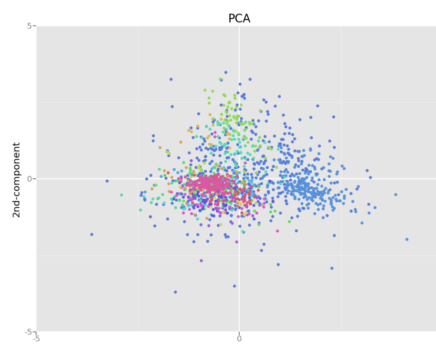

Fig. 3: Word embeddings PCA after training

While training, the word embedding vectors increase their norms, this effect is reflected in Figure 2 and Figure 3, where a PCA plot of the word representations before and after training is shown. In this graph, the 36 training tags are plotted (overlapped words appear as a separated colour). It does not seem visually clear in 2 dimensions how the data could be separated; but the further transformations of these representations and the non linearity added by the network are able to identify semantic regions, to some extent.

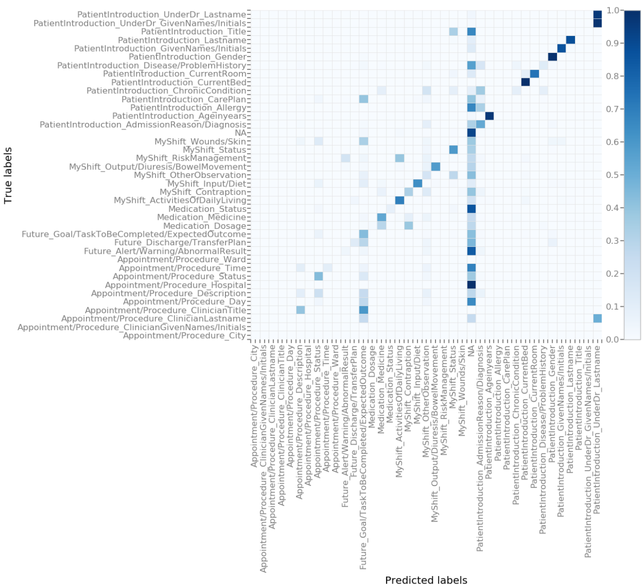

Fig. 4: Method 1 validation set confusion matrix

Figure 4 presents a confusion matrix of the neural network output when predicting on the validation set. The neural net achieves a high score for some of the easy tags, such as *PatientIntroduction_ CurrentBed*, *PatientIntroduction_ CurrentRoom*, *PatientIntroduction_ Gender*, while having a poor performance in complex tags like *Future_ Goal/ TaskToBeCompleted/ ExpectedOutcome* or *PatientIntroduction_ CarePlan*, but also in other easy tags like *PatientIntroduction_ UnderDr_ GivenNames/ Initials* (which are misclassified as *PatientIntroduction_ UnderDr_ Lastname*). As it can be seen in the plot, many samples are being wrongly classified as *"NA"*. The ensemble approach of method 2 will try to tackle this misclassification behaviour.

3.2 Method 2

In this case, the random forest uses a randomly initialized matrix intersected with the Googlenews pretrained embeddings as input features, and predicts whether the token belongs to the *"NA"* label or not. Later on, the neural network of the second step discriminates between the remaining labels. This latter model was trained using a context window of 7 for 50 epochs.

Figure 5 presents a PCA plot of the word embeddings corresponding to the tag *"NA"*, the remaining tags, and also separates word types that belong to *"NA"* and, at least, some other label.

Table 4 presents a summary of the ensemble method results. A detailed, per tag, analysis can be found in Table 8.

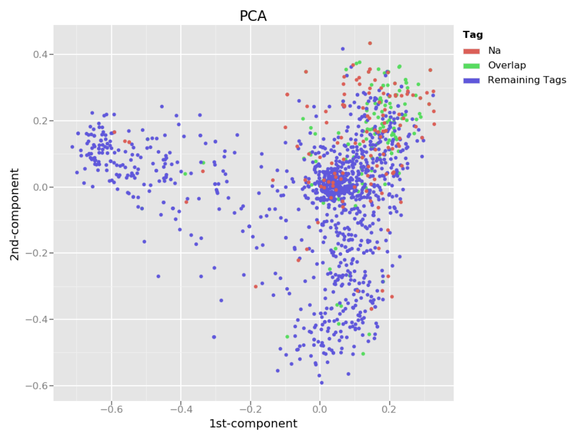

Fig. 5: PCA on random forest's word embedding inputs

Table 4: Method 2 results

Dataset	Macro average			Micro average			NA		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Training	0.768	0.699	0.718	0.81	0.861	0.835	0.859	0.791	0.824
Validation	0.434	0.397	0.385	0.541	0.546	0.543	0.846	0.835	0.84
Test	0.425	0.383	0.345	0.49	0.517	0.503	0.849	0.779	0.813

Figure 6 shows the confusion matrix when using the ensemble model to predict on the validation set. This time, the *NA* false-positive results are reduced, in comparison to the first approach. While incrementing the number of true-positives, some mistakes are translated to other categories, the most prominent being: *Appointment/ Procedure_Description*, *Future_Goal/ TaskToBeCompleted/ ExpectedOutcome*, *MyShift_OtherObservation*, and *PatientIntroduction_AdmissionReason/ Diagnosis*.

The non-ensemble method achieved a macro-averaged F1-score of 0.308 on the 35 tags (all tags excluding *"NA"*) and the ensemble system performed at 0.345. This means the ensemble method performs better overall, obtaining higher F1-results for 17 out of the 36 classes (with improvements of up to 0.3), maintaining the same scores for 8 of them, and lowering them in 11 cases (with decrements of max 0.08). Some of these improvements imply that labels that previously had an F1-score of 0.0 are now getting 0.308, like the case of *PatientIntroduction_CarePlan*. Considering the micro-averaged, these results are translated to 0.514 and 0.503, respectively. Analysing this updates, the F1-value goes down due to the decrease in the micro-averaged precision (the micro-averaged recall increases). In the vast majority of the cases, the *"NA"* false-positive classifications of method 1 are now being assigned to the remaining classes causing an increase

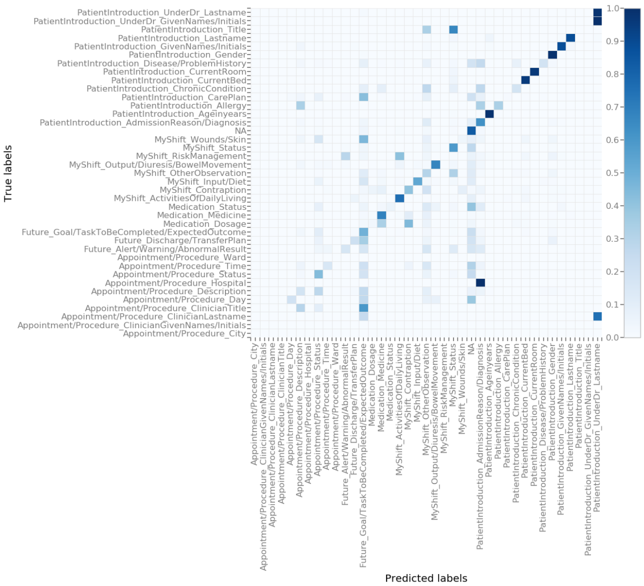

Fig. 6: Method 2 validation set confusion matrix

in their recall (lowering the false-negatives samples), and even though negatively affecting the precision, the gain is big enough to push the F1-relation up. But for some categories this statement does not hold, and it is the sum of these misclassified tokens which causes the micro-averaged reduction. Examples of this negative behaviour can be found in labels: a) *"MyShift_Input/ Diet"*, where previously correct tokens are now assigned to other classes, while also wrongly predicting tokens as members of this tag; b) *"MyShift_Status"*, where 79 false-positives are added just for gaining 2 tokens in true predictions; and c) *"PatientIntroduction_AdmissionReason/ Diagnosis"*, case in which 90 false-positives are included, with an advantage of only 3 new true-positive samples.

4 Conclusions and future work

In this report, a feed forward neural network, and a random forest - feed forward neural network ensemble method are presented as solutions for the CLEF eHealth Task 1 challenge. Both methods rely on semantic features and are domain independent (no medical features are used). While the neural network alone achieves a macro-averaged F1-score of 0.308 on the test set, considering 35 categories (all but *"NA"*), and an F1-score of 0.8 for the *"NA"* tag; the ensemble method produces better results with a macro F1-score of 0.345 on the 35 tags of the test dataset, and an F1-score of 0.813 for the *"NA"* label, increasing the precision but affecting the recall metric. This gain in F1-results suggests the second method performs better when considering the entire set of labels. On the other hand, from the micro-averaged perspective, the former method gets an F1-score of 0.514 for the 35 tags, and the latter an F1-score of 0.503, which can be explained from the relation between the general raise of the recall and the decrement in the precision.

As explained in the analysis section, the handover form presents a large number of tags, while some appear to be easy to learn, some others are clearly a complex task. One of the main problems limitations, given the Machine Learning methods implemented, is the amount of data available for training.

In this report's pipelines no pre-processing steps were applied, and the negative effects of this decision were explicitly pointed out in the analysis. There should be a prior stage in which abbreviations, misspellings, and errors during tokenization are treated.

While this work shows that semantic representations are able to help in this task, most likely, a higher performance could be achieved by incorporating features of other types; lexical features such as part-of-speech tags, or dependency parsing, as well as features resulting from external taggers. Moreover, a potentially useful characteristic, not exploited in this work, is the natural structure of the medical records, for instance, word or sentence locations, and tag-precedence.

References

1. Kelly, Liadh and Goeuriot, Lorraine and Suominen, Hanna and Nvol, Aurlie and Palotti, Joao and Zuccon, Guido. Overview of the CLEF eHealth Evaluation Lab 2016. CLEF 2016 - 7th Conference and Labs of the Evaluation Forum, Lecture Notes in Computer Science (LNCS), Springer, September, 2016.
2. Suominen, Hanna and Zhou, Liyuan and Goeuriot, Lorraine and Kelly, Liadh. Task 1 of the CLEF eHealth evaluation lab 2016: Handover information extraction. CLEF 2016 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS, September, 2016.
3. Suominen, Hanna and Zhou, Liyuan and Hanlen, Leif and Ferraro, Gabriela. Benchmarking Clinical Speech Recognition and Information Extraction: New Data, Methods, and Evaluations. JMIR Medical Informatics. April, 2015.
4. Zhou, Liyuan and Suominen, Hanna and Hanlen, Leif. Evaluation Data and Benchmarks for Cascaded Speech Recognition and Entity Extraction. ACM Multimedia 2015 Workshop on Speech, Language and Audio in Multimedia. October, 2015.
5. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR, 2013.
6. Rumelhart, D. E., Hinton, G. E., and Williams, R. J. Learning representations by back-propagating errors. Nature 323, 1986.
7. Duchi, John; Hazan, Elad; Singer, Yoram. Adaptive subgradient methods for online learning and stochastic optimization. JMLR 12, 2011.
8. Breiman, Leo. Random Forests. Machine Learning. October 1 2001.
9. Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. 2016.

5 Appendix

Table 5: Training data empirical distribution

Tag	Probability
PatientIntroduction_GivenNames/Initials	0.0140
PatientIntroduction_Lastname	0.0117
PatientIntroduction_Ageinyears	0.0290
PatientIntroduction_Gender	0.0576
PatientIntroduction_CurrentRoom	0.0064
PatientIntroduction_CurrentBed	0.0212
PatientIntroduction_UnderDr_GivenNames/Initials	0.0018
PatientIntroduction_UnderDr_Lastname	0.0213
PatientIntroduction_AdmissionReason/Diagnosis	0.0488
PatientIntroduction_Allergy	0.0016
PatientIntroduction_ChronicCondition	0.0082
PatientIntroduction_Disease/ProblemHistory	0.0173
PatientIntroduction_CarePlan	0.0042
MyShift_Status	0.0569
MyShift_Contraption	0.0052
MyShift_Input/Diet	0.0119
MyShift_Output/Diuresis/BowelMovement	0.0061
MyShift_Wounds/Skin	0.0065
MyShift_ActivitiesOfDailyLiving	0.0289
MyShift_RiskManagement	0.0014
MyShift_OtherObservation	0.0425
Appointment/Procedure_Status	0.0187
Appointment/Procedure_Description	0.0185
Appointment/Procedure_ClinicianGivenNames/Initials	0.0002
Appointment/Procedure_ClinicianLastname	0.0002
Appointment/Procedure_Day	0.0047
Appointment/Procedure_Time	0.0033
Appointment/Procedure_City	0.0002
Appointment/Procedure_Ward	0.0004
Medication_Medicine	0.0185
Medication_Dosage	0.0044
Medication_Status	0.0080
Future_Alert/Warning/AbnormalResult	0.0070
Future_Goal/TaskToBeCompleted/ExpectedOutcome	0.0584
Future_Discharge/TransferPlan	0.0105
NA	0.4443

Table 6: Validation set overlap

Tag	# Word types	Overlap #	Overlap %	Uniqueness #	Uniqueness %	Overlapping tags (top 3) (tag: %)
PatientIntroduction_Title	1	0	0	1	1	-
PatientIntroduction_GivenNames/Initials	101	7	0.069	94	0.931	PatientIntroduction_Lastname:0.714 PatientIntroduction_UnderDr_Lastname: 0.286
PatientIntroduction_Lastname	96	7	0.073	89	0.927	PatientIntroduction_GivenNames/Initials: 0.714 PatientIntroduction_UnderDr_Lastname: 0.286
PatientIntroduction_Ageinyears	51	7	0.137	44	0.863	PatientIntroduction_CurrentBed: 0.400 PatientIntroduction_Disease/ProblemHistory: 0.200 PatientIntroduction_CurrentRoom: 0.100
PatientIntroduction_Gender	9	5	0.556	4	0.444	PatientIntroduction_AdmissionReason/Diagnosis: 0.182 MyShift_OtherObservation: 0.136 Future_Goal/TaskToBeCompleted/ExpectedOutcome: 0.136
PatientIntroduction_CurrentRoom	18	17	0.944	1	0.056	PatientIntroduction_CurrentBed: 0.286 MyShift_OtherObservation: 0.143 PatientIntroduction_Disease/ProblemHistory: 0.071
PatientIntroduction_CurrentBed	25	22	0.88	3	0.12	PatientIntroduction_CurrentRoom: 0.254 MyShift_OtherObservation: 0.175 PatientIntroduction_Disease/ProblemHistory: 0.079
PatientIntroduction_UnderDr_GivenNames/Initials	29	3	0.103	26	0.897	PatientIntroduction_UnderDr_Lastname: 0.600 Appointment/Procedure_ClinicianLastname: 0.400
PatientIntroduction_UnderDr_Lastname	55	7	0.127	48	0.873	PatientIntroduction_UnderDr_GivenNames/Initials: 0.333 PatientIntroduction_GivenNames/Initials: 0.222 Appointment/Procedure_ClinicianLastname: 0.222
PatientIntroduction_AdmissionReason/Diagnosis	256	118	0.461	138	0.539	PatientIntroduction_Disease/ProblemHistory: 0.132 NA: 0.118 MyShift_OtherObservation: 0.110
PatientIntroduction_Allergy	3	1	0.333	2	0.667	PatientIntroduction_AdmissionReason/Diagnosis: 1.000
PatientIntroduction_ChronicCondition	10	7	0.7	3	0.3	PatientIntroduction_AdmissionReason/Diagnosis: 0.400 PatientIntroduction_Disease/ProblemHistory: 0.133 MyShift_OtherObservation: 0.133
PatientIntroduction_Disease/ProblemHistory	157	89	0.567	68	0.433	PatientIntroduction_AdmissionReason/Diagnosis: 0.165 NA: 0.154 Future_Goal/TaskToBeCompleted/ExpectedOutcome: 0.095
PatientIntroduction_CarePlan	91	73	0.802	18	0.198	Future_Goal/TaskToBeCompleted/ExpectedOutcome: 0.168 NA: 0.137 MyShift_OtherObservation: 0.089
MyShift_Status	76	61	0.803	15	0.197	MyShift_OtherObservation: 0.192 NA: 0.164 Future_Goal/TaskToBeCompleted/ExpectedOutcome: 0.100
MyShift_Contraption	39	18	0.462	21	0.538	PatientIntroduction_AdmissionReason/Diagnosis: 0.153 MyShift_OtherObservation: 0.136 NA: 0.119
MyShift_Input/Diet	38	22	0.579	16	0.421	Future_Goal/TaskToBeCompleted/ExpectedOutcome: 0.140 MyShift_OtherObservation: 0.110 NA: 0.100
MyShift_Output/Diuresis/BowelMovement	26	13	0.5	13	0.5	Future_Goal/TaskToBeCompleted/ExpectedOutcome: 0.138 PatientIntroduction_AdmissionReason/Diagnosis: 0.123 NA: 0.108
MyShift_Wounds/Skin	18	14	0.778	4	0.222	Future_Goal/TaskToBeCompleted/ExpectedOutcome: 0.167 NA: 0.100 PatientIntroduction_Disease/ProblemHistory: 0.083
MyShift_ActivitiesOfDailyLiving	48	20	0.417	28	0.583	NA: 0.116 PatientIntroduction_AdmissionReason/Diagnosis: 0.098 Future_Goal/TaskToBeCompleted/ExpectedOutcome: 0.098
MyShift_RiskManagement	37	21	0.568	16	0.432	NA: 0.133 PatientIntroduction_AdmissionReason/Diagnosis: 0.111 MyShift_ActivitiesOfDailyLiving: 0.100
MyShift_OtherObservation	168	110	0.655	58	0.345	NA: 0.135 MyShift_Status: 0.116 Future_Goal/TaskToBeCompleted/ExpectedOutcome: 0.107
Appointment/Procedure_Status	43	40	0.93	3	0.07	NA:0.194 Future_Goal/TaskToBeCompleted/ExpectedOutcome: 0.165 MyShift_OtherObservation: 0.082
Appointment/Procedure_Description	122	64	0.525	58	0.475	Future_Goal/TaskToBeCompleted/ExpectedOutcome:0.152 PatientIntroduction_AdmissionReason/Diagnosis: 0.129 PatientIntroduction_CarePlan: 0.103
Appointment/Procedure_ClinicianTitle	8	6	0.75	2	0.25	Future_Goal/TaskToBeCompleted/ExpectedOutcome: 0.400 Appointment/Procedure_Description: 0.200 PatientIntroduction_CarePlan: 0.200
Appointment/Procedure_ClinicianLastname	4	2	0.5	2	0.5	PatientIntroduction_UnderDr_Lastname: 0.500 PatientIntroduction_UnderDr_GivenNames/Initials: 0.500
Appointment/Procedure_Hospital	1	0	0	1	1	-
Appointment/Procedure_Day	13	12	0.923	1	0.077	NA: 0.191 PatientIntroduction_Disease/ProblemHistory: 0.106 Medication_Status: 0.085
Appointment/Procedure_Time	12	11	0.917	1	0.083	NA: 0.281 Future_Goal/TaskToBeCompleted/ExpectedOutcome: 0.188 Future_Discharge/TransferPlan: 0.094
Medication_Medicine	64	19	0.297	45	0.703	Future_Goal/TaskToBeCompleted/ExpectedOutcome: 0.145 NA: 0.129 Appointment/Procedure_Description: 0.097
Medication_Dosage	41	24	0.585	17	0.415	NA: 0.159 PatientIntroduction_Disease/ProblemHistory: 0.087 MyShift_OtherObservation: 0.087
Medication_Status	29	24	0.828	5	0.172	NA: 0.168 Future_Goal/TaskToBeCompleted/ExpectedOutcome: 0.118 PatientIntroduction_Disease/ProblemHistory: 0.092
Future_Alert/Warning/AbnormalResult	18	10	0.556	8	0.444	NA: 0.154 Future_Goal/TaskToBeCompleted/ExpectedOutcome: 0.115 PatientIntroduction_Disease/ProblemHistory: 0.096
Future_Goal/TaskToBeCompleted/ExpectedOutcome	171	138	0.807	33	0.193	NA: 0.168 PatientIntroduction_CarePlan: 0.107 MyShift_OtherObservation: 0.085
Future_Discharge/TransferPlan	63	49	0.778	14	0.222	NA: 0.174 Future_Goal/TaskToBeCompleted/ExpectedOutcome: 0.164 PatientIntroduction_CarePlan: 0.087
NA	292	162	0.555	130	0.445	Future_Goal/TaskToBeCompleted/ExpectedOutcome: 0.152 MyShift_OtherObservation: 0.097 PatientIntroduction_Disease/ProblemHistory: 0.087

Table 7: Method 1 per tag detailed scores

Tag	Training			Validation			Testing		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Appointment/Procedure_City	0	0	0	0	0	0	0	0	0
Appointment/Procedure_ClinicianGivenNames/Initials	0	0	0	0	0	0	0	0	0
Appointment/Procedure_ClinicianLastname	0	0	0	0	0	0	0	0	0
Appointment/Procedure_Day	1	0.225	0.367	0	0	0	0	0	0
Appointment/Procedure_Description	0.826	0.726	0.773	0.524	0.099	0.167	0.568	0.067	0.12
Appointment/Procedure_Status	0.903	0.642	0.75	0.274	0.44	0.338	0.204	0.127	0.156
Appointment/Procedure_Time	1	0.393	0.564	1	0.105	0.19	1	0.026	0.05
Appointment/Procedure_Ward	0	0	0	0	0	0	0	0	0
Future_Alert/Warning/AbnormalResult	0.75	0.356	0.483	0.048	0.042	0.044	0.5	0.071	0.125
Future_Discharge/TransferPlan	0.958	0.775	0.857	0.379	0.124	0.186	0.576	0.157	0.247
Future_Goal/TaskToBeCompleted/ExpectedOutcome	0.804	0.923	0.859	0.333	0.435	0.377	0.06	0.252	0.097
Medication_Dosage	1	0.054	0.103	1	0.017	0.034	0	0	0
Medication_Medicine	0.818	0.86	0.839	0.653	0.531	0.586	0.779	0.345	0.478
Medication_Status	0.94	0.691	0.797	0.75	0.086	0.154	0.889	0.055	0.104
MyShift_ActivitiesOfDailyLiving	0.975	0.963	0.969	0.559	0.698	0.621	0.869	0.695	0.772
MyShift_Contraption	0.902	0.841	0.871	0.302	0.33	0.315	0.01	0.023	0.014
MyShift_Input/Diet	0.957	0.881	0.918	0.909	0.633	0.746	0.818	0.841	0.829
MyShift_OtherObservation	0.883	0.881	0.882	0.271	0.176	0.214	0.161	0.138	0.149
MyShift_Output/Diuresis/BowelMovement	0.842	0.615	0.711	0.974	0.578	0.725	0.118	0.04	0.06
MyShift_RiskManagement	0	0	0	0	0	0	0	0	0
MyShift_Status	0.902	0.915	0.909	0.59	0.59	0.59	0.692	0.765	0.727
MyShift_Wounds/Skin	1	0.655	0.791	0	0	0	0.625	0.2	0.303
PatientIntroduction_AdmissionReason/Diagnosis	0.912	0.949	0.93	0.698	0.522	0.597	0.286	0.758	0.416
PatientIntroduction_Ageinyears	0.976	0.988	0.982	0.993	0.968	0.98	0.953	0.929	0.941
PatientIntroduction_Allergy	0	0	0	0	0	0	0	0	0
PatientIntroduction_CarePlan	1	0.056	0.105	0	0	0	0	0	0
PatientIntroduction_ChronicCondition	0.929	0.557	0.696	0.063	0.091	0.074	0	0	0
PatientIntroduction_CurrentBed	0.984	1	0.992	0.872	1	0.931	0.931	0.96	0.945
PatientIntroduction_CurrentRoom	1	1	1	1	0.741	0.851	0.99	0.98	0.985
PatientIntroduction_Disease/ProblemHistory	0.912	0.85	0.88	0.8	0.12	0.209	0.063	0.024	0.035
PatientIntroduction_Gender	0.96	0.994	0.977	0.899	0.989	0.942	0.985	0.736	0.842
PatientIntroduction_GivenNames/Initials	0.943	0.975	0.959	0.9	0.865	0.882	0.759	0.85	0.802
PatientIntroduction_Lastname	0.949	0.949	0.949	0.967	0.88	0.921	0.864	0.752	0.804
PatientIntroduction_UnderDr_GivenNames/Initials	0	0	0	0	0	0	0	0	0
PatientIntroduction_UnderDr_Lastname	0.907	0.967	0.936	0.63	0.963	0.762	0.674	0.969	0.795
NA	0.92	0.979	0.948	0.696	0.92	0.793	0.723	0.894	0.8

Table 8: Method 2 per tag detailed scores

Tag	Training			Validation			Testing		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Appointment/Procedure_City	0	0	2	0	0	0	0	0	0
Appointment/Procedure_ClinicianGivenNames/Initials	0	0	2	0	0	0	0	0	0
Appointment/Procedure_ClinicianLastname	0	0	2	0	0	0	0	0	4
Appointment/Procedure_Day	30	9	10	0.769	0.75	0.759	5	5	21
Appointment/Procedure_Description	133	48	24	0.735	0.847	0.787	40	76	182
Appointment/Procedure_Status	108	51	51	0.679	0.679	0.679	40	181	51
Appointment/Procedure_Time	16	5	12	0.762	0.571	0.653	3	5	16
Appointment/Procedure_Ward	0	0	3	0	0	0	0	0	0
Future_Alert/Warning/AbnormalResult	41	11	18	0.788	0.695	0.739	4	46	20
Future_Discharge/TransferPlan	69	12	20	0.852	0.775	0.812	19	34	70
Future_Goal/TaskToBeCompleted/ExpectedOutcome	412	226	84	0.646	0.831	0.727	188	494	198
Medication_Dosage	25	0	12	1	0.676	0.806	4	13	113
Medication_Medicine	143	44	14	0.765	0.911	0.831	121	76	56
Medication_Status	42	4	26	0.913	0.618	0.737	2	3	33
MyShift_ActivitiesOfDailyLiving	201	20	44	0.91	0.82	0.863	138	88	44
MyShift_Contraption	38	1	6	0.974	0.864	0.916	36	82	52
MyShift_Input/Diet	87	10	14	0.897	0.861	0.879	42	2	37
MyShift_OtherObservation	285	167	76	0.631	0.789	0.701	99	312	230
MyShift_Output/Diuresis/BowelMovement	44	8	8	0.846	0.846	0.846	43	14	21
MyShift_RiskManagement	8	0	4	1	0.667	0.8	0	0	94
MyShift_Status	418	98	65	0.81	0.865	0.837	172	202	121
MyShift_Wounds/Skin	42	3	13	0.933	0.764	0.84	1	15	23
PatientIntroduction_AdmissionReason/Diagnosis	369	107	45	0.775	0.891	0.829	340	205	204
PatientIntroduction_Ageinyears	243	9	3	0.964	0.988	0.976	277	9	4
PatientIntroduction_Allergy	7	0	7	1	0.5	0.667	1	1	2
PatientIntroduction_CarePlan	18	0	18	1	0.5	0.667	2	2	154
PatientIntroduction_ChronicCondition	55	9	15	0.859	0.786	0.821	2	37	9
PatientIntroduction_CurrentBed	180	9	0	0.952	1	0.976	149	33	7
PatientIntroduction_CurrentRoom	54	1	0	0.982	1	0.991	52	2	2
PatientIntroduction_Disease/ProblemHistory	111	47	36	0.703	0.755	0.728	50	13	216
PatientIntroduction_Gender	486	20	3	0.96	0.994	0.977	374	73	4
PatientIntroduction_GivenNames/Initials	116	9	3	0.928	0.975	0.951	92	20	12
PatientIntroduction_Lastname	99	7	0	0.934	1	0.966	89	7	11
PatientIntroduction_UnderDr_GivenNames/Initials	4	0	11	1	0.267	0.421	0	0	60
PatientIntroduction_UnderDr_Lastname	177	16	4	0.917	0.978	0.947	106	64	2
NA	2984	491	787	0.859	0.791	0.824	2632	480	520