NLM's System Description for the Fourth i2b2/VA Challenge

Dina Demner-Fushman[1], MD, PhD, Emilia Apostolova*[2], MS,
Rezarta Islamaj Doğan*[3], PhD, François-Michel Lang[1], MSE,
James G. Mork[1], MSc, Aurélie Névéol*[3], PhD
Sonya E. Shooshan[1], MLS, Matthew Simpson*[1], MS, Alan R. Aronson[1], PhD

[1]Lister Hill National Center for Biomedical Communications (LHNCBC)/NLM/NIH,
[2]DePaul University, Chicago, IL 60604,
[3]National Center for Biotechnology Information (NCBI)/NLM/NIH, Bethesda, MD 20894
***Primary contributors; these authors contributed equally to this work.**

Abstract

The NLM team participated in the concept extraction, assertion classification and relation extraction tasks of the Fourth i2b2 Challenge. After exploring rule-based approaches to the tasks, the team has resorted to supervised machine learning methods. The results varied by task. The concept extraction task suffered from over-training, achieving 0.60 - 0.74 F_1-score in the evaluation (compared to 0.93 cross-validation results on the training set). The performance of the assertion classifiers in the evaluation was consistent with that in training, achieving an overall 0.93 F_1-score. An SVM classifier demonstrated mixed behavior in the relation extraction task: it was over-trained for two relations with few positive examples, and the results were consistent with training for the remaining relations (achieving an overall 0.67 F_1-score on the test set.) The improvements in relation extraction due to feature selection observed in cross-validation evaluations during training were not observed in the test evaluation. The assertion and relation classifiers applied to concepts extracted by our system maintained their respective performance, but the scores were much lower than for the ground-truth based evaluation.

INTRODUCTION

The Fourth i2b2/VA (Informatics for Integrating Biology and the Bedside/Veteran Administration) Challenge[1] consisted of three Natural Language Processing (NLP) tasks: concept extraction, assertion classification and relation extraction on deidentified clinical records. The deidentified data used in the Challenge consisted of discharge summaries from both Partners HealthCare and the MIMIC II Database of the Beth Israel Deaconess Medical Center and also discharge summaries and progress notes from the University of Pittsburgh Medical Center.

The concept task consisted of extracting three types of concepts (medical problems, tests and treatments) from the clinical records. The assertion task used the results of the concept task for determining assertions made on the medical problems according to a predefined classification of assertion types. Finally, the relation task built upon the first two tasks to extract relations asserted on the medical problems, tests and treatments of the concept task.

The NLM team participated in all three tasks of the Fourth i2b2 Challenge and relied mainly on standard machine learning approaches augmented with selected knowledge-based techniques.

DATA PROCESSING

Each of the input text files was processed by converting any UTF-8 characters to their ASCII equivalents and then applying NLM's MetaMap[2] program to discover the UMLS® Metathesaurus®[3] concepts mentioned in the text. MetaMap's quick composite phrases option was used in the processing to ensure that coherent phrases with internal prepositional phrases were considered to be a single unit. This option increases mapping accuracy by associating longer spans of text with more specific Metathesaurus concepts. The correspondence between the original text and concepts was retained for later processing.

CONCEPT EXTRACTION

The goal of the concept extraction task was to automatically identify, within clinical records, complete phrases representing medical problems, treatments, and tests. The provided training dataset consisted of 349 clinical documents for which there were manually assigned annotations (43% *problems*, 31% *treatments* and 26% *tests*).

Our approach to this task relied on supervised machine learning and features extracted from the training dataset.

First, we processed the clinical records to remove uninformative lines and identify section headings. We utilized regular expression patterns for each of these tasks. For example, a line was removed if it contained only numbers and punctuation, and a line was identified as a section heading if it ended with a colon.

We used MetaMap to parse the clinical documents and identify UMLS concepts. The noun phrases identified by MetaMap were post-processed with an ad-hoc set of rules to (1) remove leading assertion modifiers, (2) prune prepositional phrases unless they indicate a body part or location, (3) combine phrases connected by conjunctions and other syntax that denote lists, and (4) combine phrases in apposition if the appositive is set off by parentheses. We used a closed class of assertion modifiers (*no*, *possible*, *likely*, etc.), and body parts were identified based on the UMLS semantic types of the candidate mappings.

Finally, we extracted the following eight features for each processed phrase for use in training:

1. *Section heading*: The heading of the section containing the phrase. Section headings were manually normalized to be one of nine labels: test, problem, diagnosis, course, history, exam, ignore, treatment, or allergy.

2. *Tokens within phrase*: An unordered list of tokens comprising the phrase. All dates, times, and measurements identified with regular expressions were normalized to the strings <*date*>, <*time*>, and <*measurement*>.

3. *Preceding bigram*: The normalized (as in 2) two-word token group preceding the phrase.

4. *Following bigram*: The normalized (as in 2) two-word token group following the phrase.

5. *Lexical category*: An unordered list of lexical categories (i.e., parts of speech) produced by the MetaMap parser for tokens within the phrase.

6. *Grammatical fun*ction: An unordered list of MetaMap syntax types (the roles the constituents play in the constructions they belong to, for example, *head*, *modifier*) produced by the MetaMap parser for tokens within the phrase.

7. *CUIs*: An unordered list of candidate UMLS Concept Unique Identifiers (CUIs) to which the phrase potentially maps.

8. *Semantic types*: An unordered list containing the UMLS semantic type of all the phrase's candidate concept mappings.

We trained two one-against-one multi-class Support Vector Machines (SVMs)[4] for classifying each processed phrase as a *problem*, *treatment*, *test*, or *none*. In training the first SVM, we labeled each processed phrase according to an exact phrase boundary matching of the annotated training data, and for the second SVM, we used inexact matching to label the processed phrases. We used the radial basis function (RBF) kernel with parameters $C = 32.0$ and $\gamma = 0.0078125$, which were determined by a grid search using cross-validation. The 5-fold cross-validation accuracy for both multi-class SVMs was 0.93 on our training dataset.

Concept Extraction Results

Table 1 shows the results of the exact span classifier in identifying medical concepts.

Concept	**Evaluation**					
	Exact			Inexact		
	P	*R*	F_1	*P*	*R*	F_1
Problem	0.88	0.39	0.53	0.95	0.42	0.58
Treatment	0.88	0.46	0.60	0.95	0.49	0.65
Test	0.90	0.47	0.62	0.96	0.50	0.66
Overall	*0.88*	*0.43*	*0.58*	*0.96*	*0.46*	*0.63*

Table 1. Results for the concept extraction task using an exact span classifier (*P* = Precision, *R* = Recall, F_1 = F_1-score).

Concept	**Evaluation**					
	Exact			Inexact		
	P	*R*	F_1	*P*	*R*	F_1
Problem	0.67	0.50	0.57	0.88	0.65	0.75
Treatment	0.76	0.52	0.62	0.92	0.62	0.74
Test	0.75	0.52	0.62	0.92	0.63	0.75
Overall	*0.72*	*0.51*	*0.60*	*0.90*	*0.63*	*0.74*

Table 2. Results for the concept extraction task using an inexact span classifier (*P* = Precision, *R* = Recall, F_1 = F_1-score).

Table 2 shows the results of the inexact span classifier. Under both evaluation metrics, the inexact span classifier improves recall compared to that of the exact span classifier, but also somewhat reduces precision. The inexact span classifier achieved better F_1-scores under both the exact span evaluation metric (0.60 compared to 0.58 of the exact classifier) and the inexact span metric (0.74 compared to 0.63.)

Both our exact and inexact SVM classifiers performed poorer than expected on the test data. Given their 0.93 cross-validation accuracy, the classifiers' performance is likely the result of an over fitting of the training data.

The low performance on the exact span metric demonstrates our difficulty in reproducing the annotated phrase boundaries. This can be explained by noting inconsistencies in the training dataset and the degree to which parses produced by MetaMap differed from the annotated examples. In retrospect, a better approach may have been to first extract the medical concepts (e.g., with MetaMap) and a set of concept-specific features for use in training, and then use a rule-based approach for expanding the classified concepts to the desired phrase boundaries.

ASSERTION CLASSIFICATION

The assertion classification task involved classifying medical problems into one of six categories shown in Table 3. Figure 1 summarizes the distribution of the 6 assertion categories in the training dataset.

Assertion Category	Example
Present	*He has pneumonia*
Absent	*No pneumonia was suspected*
Hypothetical	*Return to the emergency room if you develop fevers*
Possible	*Pneumonia is possible / probable*
Conditional	*Penicillin causes a rash*
Associated with someone else	*Brother had asthma*

Table 3. Assertion categories and examples.

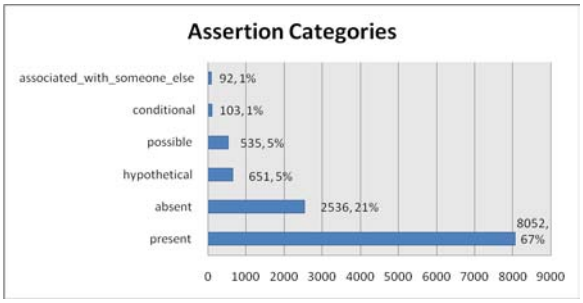

Figure 1. Distribution of the assertion categories across the 11,969 medical problems in the training dataset.

Since *present* is the prevalent assertion category (see Figure 1), a naïve baseline created by assigning all medical problems to this category would result in an F_1-score of 0.67.

Rule-based Approach

We first evaluated the performance of a rule-based algorithm developed for identifying contextual features from clinical text – ConText[5]. The algorithm relies on hand-crafted sets of trigger terms in proximity of clinical conditions to discover if the conditions are affirmed, negated, or possible; recent, historical, or hypothetical; experienced by the patient or other. We slightly modified the algorithm by extending the list of *possible* trigger-terms and introducing a small set of *conditional* trigger terms. We also disregarded *historical* cues as the challenge task does not differentiate between historical or recent medical problems. Table 4 shows that results are considerably better than the naïve baseline.

Assertion Category	TP	FN	FP	R	P	F_1
Present	7344	708	817	0.91	0.89	0.91
Absent	1755	781	271	0.69	0.87	0.77
Hypothetical	469	182	29	0.72	0.94	0.82
Possible	277	258	741	0.52	0.27	0.36
Conditional	36	67	143	0.35	0.20	0.26
N/A	52	9	40	0.57	0.85	0.68
Overall	9933	2036	2010	0.83	0.83	0.83

Table 4. Results for the assertion classification task using the modified ConText algorithm on the training dataset. (TP=True Positive, FN=False Negative, FP=False Positive, R=Recall, P=Precision, F_1= F_1-score)

Machine Learning Approach

As fine-tuning the ConText trigger terms and introducing new rules proved laborious, we next explored supervised machine learning. We modeled the problem as a classification task that assigns each medical problem into one of the six categories. We trained a one-against-all SVM[4] classifier (a series of binary classifiers for each assertion category against all other categories). Empirically, we identified an optimal set of features described below (the GATE[6] framework was used to generate and experiment with feature sets).

Feature Set:

1. *Token window of size 5*: Tokens surrounding the medical problem (within sentence boundaries). Numbers were normalized (converted to the string *$number*). Tokens belonging to concepts were converted to their corresponding concept types (e.g.

coronary bypass surgery was substituted by the concept's category – *treatment*).

2. *Negative prefix*: This feature targets the discovery of *absent* medical problems identified as such by the presence of a prefix, as in *a-febrile* or *non-tender*. Possible values are *a-*, *ab-*, *un-*, *an-*, *anti-*, *dis-*, *non-*, *in-*, *il-*, *ir-*, or *im-*.

3. *Section heading preceding the problem concept*: Section headings could be helpful in identifying most assertion categories. For example, problems that fall under the heading *Family History* typically fall into the *not associated with patient* category. Headings were identified as the last string preceding the problem concept that matches the regular expression "*Beginning of line, One or more characters, Colon, White space, End of line*".

4. *ConText Cues*: Occasionally, *cues* and *trigger terms* outside the limitations of the 5-token window were necessary for a human reader to identify the assertion category. This feature was used to identify ConText cues preceding or following the medical concept outside the token window size.

5. *Semantic Type*: *Conditional* medical problems occur under certain circumstances (such as, allergies). For example, *penicillin causes a rash, dyspnea on exertion*. A dictionary was manually compiled to map problem concepts to such semantic types.

In addition, token-window Part-of-Speech, UMLS term and semantic type features were also considered. However, these features had no positive effect on the system performance and were excluded from the final submission run.

Overall, SVM showed considerable performance improvement over our rule-based approach. An F_1-score of 0.94 was achieved on a held-out portion of the training set. This result was consistent with the performance of the SVM classifier in the test run described below.

Assertion Classification Results

We submitted two system runs on the test dataset of 477 clinical records. In both cases we used the SVM classifier trained on the ground truth concepts from the test dataset (linear kernel function, cost 0.7, uneven margins parameter 0.5).

Table 5 shows the results of classifying the ground truth concepts provided with the test data. For both the SVM and the rule-based approaches, the categories most difficult to identify were *conditional* and *possible* medical problems. The difficulty with these two categories could stem from higher vocabulary variability, as well as information outside the 5-token window disregarded by our model. The small number of *conditional* statements in the training data was also problematic.

To evaluate the overall system performance, we also submitted an assertion run on the concepts identified by our system (See Table 6). In identifying the exact span problem concepts, our system achieved an F_1-score of 0.58. The assertion classifier achieved an F_1-score of 0.52 decreasing the overall system F_1-score by about 10% (comparable to the F_1-score of 0.93 on the ground truth concepts). The difference of the relative performance across assertion categories in the two runs is due to the varying performance of identifying concepts of different assertion categories. For example, in the concept extraction task our system achieved an F_1-score of 0.52 on identifying the *present* exact span problem concepts, and an F_1-score of 0.69 for the *absent* exact span problems.

Assertion Category	TP	FN	FP	R	P	F_1
Present	1280	217	930	0.98	0.93	0.96
Absent	3331	278	145	0.92	0.95	0.94
Hypothetical	568	149	38	0.79	0.94	0.86
Possible	449	434	102	0.51	0.81	0.63
Conditional	44	127	14	0.26	0.76	0.38
N/A	111	34	10	0.77	0.92	0.83
Overall	17311	1239	1239	0.93	0.93	0.93

Table 5. Results for the assertion classification task using the SVM classifier and the ground truth concepts. (N/A –associates with someone else)

Assertion Category	TP	FN	FP	R	P	F_1
Present	6028	699	493	0.46	0.55	0.50
Absent	2065	154	595	0.57	0.77	0.66
Hypothetical	266	451	126	0.37	0.68	0.48
Possible	195	688	130	0.22	0.60	0.32
Conditional	28	143	14	0.16	0.67	0.26
N/A	59	86	35	0.41	0.63	0.49
Overall	8641	9909	5835	0.47	0.59	0.52

Table 6. Results for the assertion classification task using the SVM classifier and concepts identified by our system (N/A –associates with someone else)

RELATION EXTRACTION

The relation extraction task involved identifying eight relationships between *problem*, *treatment*, and *test* concepts occurring in the same sentence and three implicit *absent* relations shown in Table 7.

Figure 2 summarizes the distribution of the relation categories in the training dataset.

Relations	Example
Problem Interacts with Problem (**PIP**)	Chronic diarrhea with multiple admissions for dehydration
No Problem/Problem Relation (*noPIP*)	No masses or thrombi are seen in the left ventricle
Test Reveals Problem (**TeRP**)	Chest x-ray revealed mild pulmonary edema
Test Conducted for Problem (**TeCP**)	She eventually underwent a cardiac cath to assess coronary disease
No Test/Problem relation (*noTeP*)	Had subsequent abnormal stress test and pefusion imaging showed EF 38% with perfusion defects .
Treatment is Administered for Problem (**TrAP**)	For her depression , continued on Citalopram 10 mg daily
Treatment is *not* Administered for Problem (**TrNAP**)	She was initially started on Bactrim , but this was stopped as she was afebrile
Treatment Causes Problem (**TrCP**)	Left ankle was notable for swelling secondary to status post left ankle surgery
Treatment Improves Problem (**TrIP**)	Her chest pain was controlled with morphine.
Treatment Worsens Problem (**TrWP**)	fevers recurred on this antibiotic regimen
No Treatment/ Problem relation (*noTrP*)	Past Medical History: Arthritis , h/o Bell's Palsy , HOH , s/p Tonsillectomy

Table 7. Relation categories and examples (problem concepts are shown in orange, treatments in blue, tests in green).

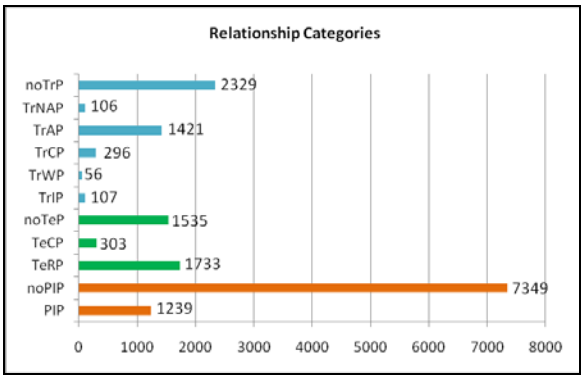

Figure 2. Distribution of the relation categories across the 5,261 relations in the training dataset, and additional 11,213 relation candidates automatically generated as "negative" examples.

Extracting relation candidates

As the task involved extracting relationships bounded by sentences, only a limited number of relation candidates could be generated. For each pair of (*problem*, *problem*) concepts, there was one PIP candidate, for each pair of (*problem*, *test*) concepts, there were two relation candidates TeCP and TeRP, for each pair of (*problem*, *treatment*) concepts, there were five relation candidates TrAP, TrNAP, TrIP, TrWP and TrCP. The combinatorial candidate generation was overly productive: only 14% of the candidate relations were in the ground truth.

Relation Extraction Approach

The prevalent relation categories are *noPIP*, *TeRP* and *noTrP*. A naïve baseline created by assigning each eligible candidate to the *TeRP* category resulted in an F_1-score of 0.39. As this baseline was comparable to our exploration of a rule-based approach, we turned to machine learning, and trained an in-house implementation of a linear SVM[7] for each relationship, considering the following features within sentence boundaries:

1. Bag-of-Words (BOW)
2. BOW before the 1st concept
3. BOW after the 2nd concept
4. BOW between two concepts
5. UMLS CUIs (obtained using MetaMap)
6. Semantic Types (STs)
7. Assertion value (for problem concepts)
8. Co-occurrence of concept pair in MEDLINE®
9. Co-occurrence of concept pair in our database of treatment/problem pairs[8].

In practice, UMLS mappings (CUIs and STs) were available for only 70% of the concepts, but they proved to be useful for classification. On the other hand, co-occurrence data gathered either from MEDLINE or other sources was too sparse to be useful.

In addition to using all features for each classifier, we used an iterative feature selection procedure. Features that were assigned the lowest weight by the classifier during the learning phase were removed, and the classifier was re-trained on the remaining features. This iterative process was repeated until no further improvement was observed. Performance evaluation on the training set was carried out using 5-fold cross-validation.

Relation Extraction Results

We submitted three system runs on the test dataset of 477 clinical records. For all runs we used the SVM classifier trained on the ground truth concepts from the training set of 349 clinical documents. Run 1 (Table 8) used the classifiers without feature selection, applied on the ground truth concepts and assertions.

Relation	TP	FN	FP	R	P	F_1
TrIP	73	125	42	0.37	0.63	0.47
TrWP	18	125	34	0.13	0.35	0.18
TrCP	218	226	189	0.49	0.54	0.51
TrAP	1771	716	854	0.71	0.67	0.69
TrNAP	38	153	63	0.20	0.38	0.26
PIP	1059	927	541	0.53	0.66	0.59
TeRP	2222	811	327	0.73	0.87	0.79
TeCP	231	357	158	0.39	0.59	0.47
Overall	5630	3440	2208	0.62	0.72	0.67

Table 8. Submission Run 1. Results from SVM classifier using ground truth concepts and assertions (Concept With Matching Span and Relation)

As observed on the training set, the best results were obtained for the most frequent relations, TeRP and TrAP. However, for the less frequent relations, TrWP and TrNAP, the performance was very low.

Run 2 (Table 9) also used the classifiers without feature selection, applied to the concepts and assertions supplied by our system. While the overall performance is significantly lower, reflecting the mismatch between the extracted concepts and the ground truth concepts, the trend is similar to that observed in Run 1, with the best performance for TeRP and TrAP, worst for TrWP.

Relation	TP	FN	FP	R	P	F_1
TrIP	19	179	34	0.10	0.36	0.15
TrWP	2	141	18	0.01	0.10	0.03
TrCP	31	413	151	0.07	0.17	0.10
TrAP	432	2055	930	0.17	0.32	0.22
TrNAP	13	178	30	0.07	0.30	0.11
PIP	275	1711	1161	0.14	0.19	0.16
TeRP	616	2417	728	0.20	0.46	0.28
TeCP	76	512	149	0.13	0.34	0.18
Overall	1464	7606	3201	0.16	0.31	0.21

Table 9. Submission Run 2. Results from SVM classifier using extracted concepts and assertions (Concept With Matching Span and Relation)

The results for Run 3 that used the classifiers with feature selection applied to the ground truth concepts were very similar to those of Run 1. The significant improvement due to feature selection (obtained in the 5-fold cross validation on the training set) was not observed in the test evaluation.

References

1. Fourth i2b2/VA Shared-Task and Workshop: Challenges in Natural Language Processing for Clinical Data. https://www.i2b2.org/NLP/Relations/Main.php

2. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc. 2010 May 1;17(3):229-36.

3. Lindberg DA, Humphreys BL, McCray AT. The unified medical language system. Methods Inf Med 1993;32(4):281–91.

4. Chang CC, Lin CJ. LIBSVM: a library for support vector machines; 2001. Software at http://www.csie.ntu.edu.tw/_cjlin/libsvm.

5. Chapman WW, Chu D, Dowling JN. ConText: An algorithm for identifying contextual features from clinical text. Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing. 2007;81-88.

6. Cunningham H, Maynard D., Bontcheva K, Tablan V. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). 2002.

7. Vapnik V, Statistical Learning Theory. New York: John Wiley & Sons Inc., 1998.

8. Névéol A, Lu Z. Automatic integration of drug indications from multiple health resources. 1st ACM International Health Informatics Symposium.