WVU NLP Class Participation in ShARe/CLEF Challenge

V. Jagannathan[1,2], D. Ganesan[1], A. Jagannathan[1], R. Kavi[1], A. Lamb[1], F. Peters[1], S. Seeger[1]

[1]West Virginia University, [2]M*Modal

Abstract. The spring 2013 graduate class in NLP decided to participate in the ShARe/CLEF challenge Tasks 1 & 2. The timing for the challenge coincided nicely with the spring semester session. There were six students in the graduate class, and the challenge tasks appeared to be good material to expose students to a practical task faced by healthcare industry. The general approach used by the class is to use CRF learning algorithm using Factorie – a scala-language based toolkit. The F-scores for best results for Task 1a relaxed – 0.801, 1a strict – 0.554, 1b relaxed – 0.625 & 1b strict 0.349, respectively. Task 2 was attempted as a group task. F-scores were 0.426 and 0.428 for strict and relaxed respectively. It was a real challenge to focus on the challenge as a class project. The students did learn how to apply NER CRF engine to a practical problem.

1 Introduction

Natural Language Processing (NLP) challenges have been a long standing tradition in both the main stream computer science discipline and the medical informatics community. Understanding clinical text and providing practical solutions based on the information extracted from such clinical text is currently a burgeoning industry. The solutions range from supporting revenue cycle tasks, such as computer-assisted coding to extracting quality measures to providing clinical decision support.

There has been an avalanche of work in the NLP domain by the medical informatics research community. And, a number of research challenges, such as the ones hosted by the i2b2 community. The fifth such challenge focused on the NLP task of co-reference resolution [1]. These challenges in academia serve an important purpose of furthering the state of the art and practice of NLP and equipping the next generation of students to address practical solutions in the marketplace. With this goal in mind, we engaged the entire class, albeit a small one, to focus their energies on participating in the ShARe/CLEF NLP challenge.

2 Approach

The challenge task involves recognizing problem concepts in clinical text and encoding them using SNOMED. In the NLP literature the task of recognizing and annotating concepts is described as a "named entity recognition (NER)" task. The same

was broken into two tasks: Task 1 and Task 2, where task 1 focused on recognizing problem concepts and encoding them and Task 2 involved recognizing acronyms and encoding them.

The various broad approaches to addressing NLP processing can be characterized as follows: rules-based, machine learning approaches or hybrid. Rule-based approaches rely on recognizing word patterns – typically using "regex" pattern matchers. Of recent, there has been an impressive array of results that show machine-learning approaches can be effective for the tasks at hand. Hybrid approaches take the practical aspect of both approaches to provide a combined solution.

For the class, we decided to explore machine learning approaches, and in particular to use the Conditional Random Field (CRF) approach. CRF has been used successfully in a number of research efforts. We also decided to use the Scala-based toolkit Factorie [2].

Figure 1 gives the general pipeline all the students used in creating the solution.

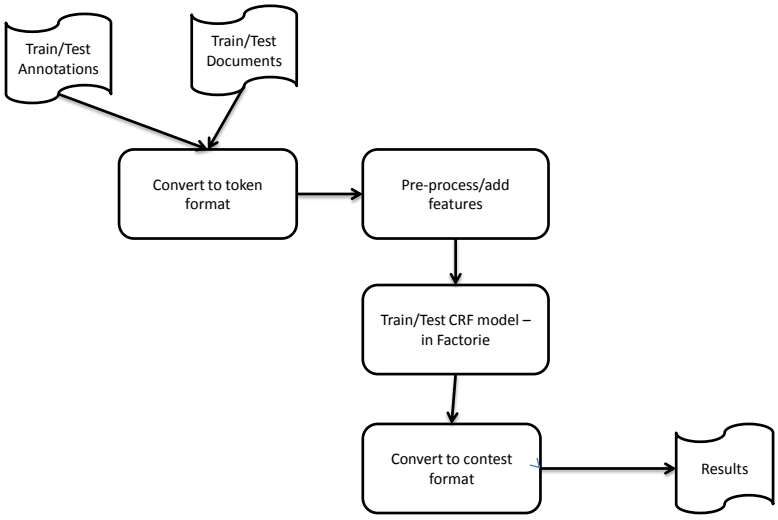

Figure 1: Approach used

The instructor provided the basic conversion routines for the class. The Factorie toolkit utilized an in-line token format, while the contest data was an off-line annotation. So, a sentence such as this:

```
The patient is a 40-year-old female with complaints of headache and
dizziness.
```

Along with this annotation:

```
00098-016139-
DISCHARGE_SUMMARY.txt||Disease_Disorder||C0018681||330||338
```

```
  00098-016139-
DISCHARGE_SUMMARY.txt||Disease_Disorder||C0012833||343||352
```

Results in the following token format:

```
The  X X O 00098-016139-DISCHARGE_SUMMARY.data 275 278
patient  X X O 00098-016139-DISCHARGE_SUMMARY.data 279 286
is  X X O 00098-016139-DISCHARGE_SUMMARY.data 287 289
a  X X O 00098-016139-DISCHARGE_SUMMARY.data 290 291
40-year-old  X X O 00098-016139-DISCHARGE_SUMMARY.data 292 303
female  X X O 00098-016139-DISCHARGE_SUMMARY.data 304 310
with  X X O 00098-016139-DISCHARGE_SUMMARY.data 311 315
complaints  X X O 00098-016139-DISCHARGE_SUMMARY.data 316 326
of  X X O 00098-016139-DISCHARGE_SUMMARY.data 327 329
headache         X     X     I_Disease_Disorder     00098-016139-
DISCHARGE_SUMMARY.data 330 338
and  X X O 00098-016139-DISCHARGE_SUMMARY.data 339 342
dizziness         X     X     I_Disease_Disorder     00098-016139-
DISCHARGE_SUMMARY.data 343 352
.  X X O 00098-016139-DISCHARGE_SUMMARY.data 352 352
```

The fields of the format are: the word itself, place holder for parts-of-speech tag, place holder for chunking tag, the NER-label (one of Disease-Disorder, or O for other), data file name, character offset to reconstruct the results set.

2.1 Preprocess and features

Each student tried to use various ways to add features. A few used the OpenNLP tool-kit to add parts-of-speech and chunking (noun phrase, verb phrase etc). Standard lexical features everyone tried included: suffix, prefix, word shape, punctuation and capitalization of tokens. Few tried n-grams features, looking at two tokens before and two tokens after the current token. Another feature tried by one student was a compilation of list of words from SNOMED Core. This list ignored words of length four or less and also a number of stop words.

2.2 Machine Learner – CRF

Linear chain CRF is particularly suited to address NER problems. The Factorie toolkit came prepackaged with examples of using it. As a first exercise in the semester, before the current NLP challenge data set was released, the students practiced with data set from i2b2 challenge that provided annotations of problems, medications and tests. They continued to enhance the CRF engine when the challenge data was released. The parameter estimation and training the CRF-model was done using Factorie's default engine – which used a stochastic gradient descent algorithm. One student experimented with Gibbs sampling. Factorie provides a variety of approaches to optimize and select parameters for machine learning the model that best predicts the label associated with a specific token. The output of the machine learner is simply an assignment of a label to the token. In the case of the challenge, we only had two pos-

sible outputs: I_Disease_Disorder or O (for other). This was a simple in-out encoding of a token.

2.3 Post-processing the output

The basic post-processing needed here was to concatenate consecutive I_Disease_Disorder tokens to single NER span. One student attempted to find discontinuous NER spans representing a coherent concept using rules, while all others avoided addressing such spans. The last step of this process was to take the token format and generate the pipe-delimited format of annotations used by the challenge organizers.

2.4 SNOMED CUI Mapping

By and large, all the students landed up trying to use Levenshtein distance to SNOMED CUI descriptions to NER label found in the previous step. Students experimented with various distances from 1 to 4. One student tried using Apache Lucene search on the NER concept over the SNOMED CUI descriptions. The process, however was excruciatingly slow – as there are lots of NER labels in the data set and each one needs to be compared with thousands of SNOMED concepts and took for all students many hours of computation time. One student used Scala parallelization and successfully sped up the computation time.

2.5 Acronym expansion and mapping

The second task of the challenge which involved recognizing medical abbreviations and mapping them to SNOMED CUIs was tackled as a group class project, as opposed to individual effort. Few students concentrated their efforts in compiling a dictionary of acronyms. We had multiple sources of acronyms to work with: 1) training data, 2) UMLS data sets, 3) general web sites. In addition, tricks to generate acronyms from general descriptions was also attempted – such as taking a UMLS description of “Congestive Heart Failure” and generating “CHF” from it. Obviously, it will have lots of noise, but those generated acronyms are unlikely to occur in real text. Some manual pruning of this set was also attempted. The persons who had the highest scores on NER recognition in training data were entrusted to run the machine learner with the acronym lexicon as features. The acronyms output as NER labels were then matched to SNOMED CUIs as before.

3 Results discussions

The overall results obtained by the class are as follows: F-scores for best results for Task 1a relaxed – 0.801, 1a strict – 0.554, 1b relaxed – 0.625 & 1b strict 0.349. For Task 2, F-scores were 0.426 and 0.428 for strict and relaxed respectively.

The NER labeling task, using the relaxed scoring, was the best score achieved. This is actually to be expected, as none of the students attempted to code discontinuous spans which require building a collection of post-coordination rules or more sophisticated factor graph models in CRF. The SNOMED CUI mapping was interesting in that the students failed to do well here. CRF models are ideal for recognizing and labeling problems (or medications or labs for that matter) – but once you have a label, coding it is its own domain of problem. The solutions attempted were simplistic and time consuming. Here one could approach with a rule-based solution – which potentially obviates the need for the machine learner in the first place!

4 Conclusions

One general observation of using the challenge as a class project was that it was simply too much work to be done within the constraints of one class. The students were new to both Scala and Factorie toolkit and though some had exposure to machine learning techniques, they were there to learn NLP techniques. The trade-offs that the class had to make was one really large complex project versus a range of small projects covering various topics – that are typical of such classes. None the less, it was a learning experience and the complexities of dealing with real problems were self-evident to the students.

5 Acknowledgements

We are grateful to the organizers of this challenge to give us the data set for use in a graduate class.

6 References

1. Özlem Uzuner, Andreea Bodnari, Shuying Shen, Tyler Forbush, John Pestian, Brett R. South: Evaluating the state of the art in coreference resolution for electronic medical records. JAMIA 19(5): 786-791 (2012)

2. McCallum A, Schultz K, Singh S. FACTORIE: probabilistic programming via imperatively defined factor graphs. In: Bengio Y, Schuurmans D, Lafferty J, et al, eds. Advances in Neural Information Processing Systems. Cambridge, MA, USA: MIT Press, 2009;22:1249-57.