Surgical and Pathology reports: a Unified Lexical Approach

Ying Ou (yiou6374@uni.sydney.edu.au)
School of Information Technologies
FACULTY OF ENGINEERING & INFORMATION TECHNOLOGIES

1. AIMS OF THE RESEARCH PROJECT

The main objective is to design an information extraction engine to identify and categorize named entities from narrative clinical reports, and identify co-referential relations among them.

2. INTRODUCTION

To make good use of clinical documents, not only the automatic extraction of medical named entities is required, but also extraction of the relations among these entities is needed. One of the most important relations is co-reference. Resolving co-reference among medical named entities can provide a better understanding of these entities, and facilitate further analysis of the coherence of clinical documents.

Extraction Definition

There are 5 classes of concepts needed to be extracted:

a) **Problems**: phrases that contain observations made by patients or clinicians about the patient's body or mind that are thought to be abnormal or caused by a disease, such as "anxiety", "the wound", etc.

b) **Treatments**: phrases that describe procedures, interventions, and substances given to a patient in an effort to resolve a medical problem, for example, "800 mg ibuprofen", "tube removal", etc.

c) **Tests**: phrases that describe procedures, panels, and measures that are done to a patient or a body fluid or sample in order to discover, rule out, or find more information about a medical problem, for instance, "WBC", "Blood pressure", etc .

d) **Persons**: mentions of persons or groups of people, including proper names, personal pronouns, possessive pronouns, job titles, and groups, such as "Mrs. Smith", "She", etc.

e) **Co-reference Pronouns**: pronouns that are not included in the person class, and could refer to Problem, Test, or Treatment, for example, "which", "it", etc .

3. MATERIALS

In total 309 de-identified clinical reports (i2b2_Beth corpus: 115 documents, i2b2_Partners corpus: 136 documents, ODIE_CLINICAL corpus: 28 documents, ODIE_PATH corpus: 30 documents) were used for training and evaluation drawn from the 2011 i2b2/VA Challenge.

4. SYSTEM ARCHITECTURE

Two classifiers are used in this system: CRF and SVM. Figure 1 demonstrates the detailed system architecture.

There are several components in the system:

1. **Manual annotation**: the ODIE concept annotations are unified to align with the i2b2 annotations by categorizing the concepts to the same classes defined by the i2b2 guidelines.
2. **Self-validation**: ground truth concepts from the i2b2 corpus are validated with a 100% train and test strategy and corrected if any error is found.
3. **Pre-processing**: each ODIE record is split into sentences, and then each sentence is split into tokens.

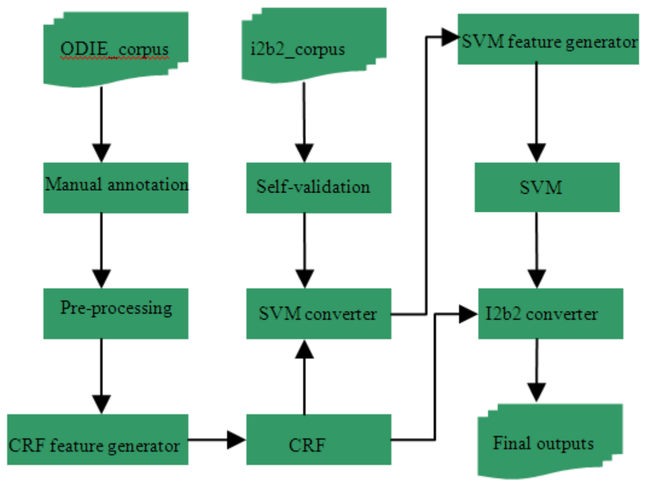

Figure 1. *System architecture*.

4. **CRF feature generator**: 5 feature sets were prepared to train the CRF, including:
- ❖Context features: bag of words with window size of three words.
- ❖Semantic features: gazetteers, lexical resources (UMLS, MOBY, SNOMED-CT), class dictionaries.
- ❖Lexical features: lemma, part of speech, chunk; expansions of abbreviations and acronyms; correction of misspelling words; number tag; lowercase form of words.
- ❖Grammatical features: title case tag.
- ❖Other features: the medication extraction system's results.

5. **CRF**: using the results from the previous component to recognize 5 entity classes.
6. **SVM converter**: create potential co-referent pairs (problem-problem, problem-pronoun, treatment-treatment, treatment-pronoun, test-test, test-pronoun, person-person, and pronoun-pronoun).
7. **SVM feature generator**: uses six feature sets to resolve co-references, which are:
- ➢Context features: three words before and after the first concept, three words before and after the second concept, words inside of each concept, the section heading of the first concept, the section heading of the second concept.
- ➢Semantic features: classes of each concept.
- ➢Position features: sentence distance apart.
- ➢Lexical features: full string match, substring match.
- ➢Grammatical features: appositive structure, head noun, number.
- ➢Domain knowledge features: Concept ID, fully specified name (FSN), preferred term, SNOMED_ID, FINDING SITE.

8. **SVM**: an SVM classifier is used to classify the co-referential relationships between each pair.
9. **I2b2 converter**: the i2b2 converter generated the outputs (concept files and chains files) according to the format required in the challenge for evaluation.

5. RESULTS

Table 1 and Table 2 illustrate the performance of the Concept Annotation experiments for exact matches of the ODIE_CLINICAL and ODIE_PATH corpora respectively.

The best results of Co-reference Resolution on the i2b2_Beth corpus and i2b2_Partners corpus are displayed in Table 3 and Table 4 respectively (using 4 evaluation metrics: BCUBED, MUC, BLANC and CEAF; P: Precision, R: Recall, F: F-score).

Table 5 and Table 6 demonstrate the End-to-end Co-reference evaluations on the ODIE_CLINICAL corpus and ODIE_PATH corpus respectively (using 2 evaluation metrics: BCUBED and CEAF; P: Precision, R: Recall, F: F-score).

Entity type	Number	Precision	Recall	F-score
Person	779	94.70%	91.78%	93.22%
Problem	535	82.82%	80.19%	81.48%
Pronoun	95	82.22%	77.89%	80.00%
Test	153	76.26%	69.28%	72.60%
Treatment	269	80.17%	70.63%	75.10%
Overall	1831	87.06%	82.69%	84.82%

Table 1. *System performance of Concept Annotation on ODIE_CLINICAL corpus*

Entity type	Number	Precision	Recall	F-score
Person	1	0	0	0
Problem	173	90.53%	88.44%	89.47%
Pronoun	0	0	0	0
Test	38	82.61%	50.00%	62.30%
Treatment	40	96.77%	75.00%	84.51%
Overall	252	90.58%	80.16%	85.05%

Table 2. *System performance of Concept Annotation on ODIE_PATH corpus*

Coref type	BCUBED			MUC			BLANC			CEAF		
	P	R	F	P	R	F	P	R	F	P	R	F
Coref test	0.985	0.962	0.973	0.292	0.638	0.401	0.745	0.608	0.65	0.925	0.965	0.945
Coref person	0.879	0.843	0.861	0.682	0.932	0.788	0.927	0.556	0.598	0.377	0.795	0.512
Coref problem	0.968	0.908	0.937	0.418	0.743	0.535	0.811	0.656	0.708	0.773	0.901	0.832
Coref treatment	0.971	0.911	0.94	0.362	0.665	0.469	0.797	0.643	0.693	0.787	0.891	0.836
Overall	0.966	0.928	0.947	0.554	0.852	0.672	0.907	0.561	0.606	0.758	0.910	0.827

Table 3. *System scores for Co-reference Resolution on i2b2_Beth corpus*

Coref type	BCUBED			MUC			BLANC			CEAF		
	P	R	F	P	R	F	P	R	F	P	R	F
Coref test	0.985	0.946	0.965	0.234	0.696	0.351	0.663	0.589	0.616	0.911	0.967	0.938
Coref person	0.93	0.913	0.921	0.902	0.904	0.903	0.861	0.941	0.897	0.459	0.466	0.463
Coref problem	0.989	0.893	0.939	0.097	0.674	0.170	0.702	0.526	0.546	0.735	0.925	0.819
Coref treatment	0.993	0.882	0.934	0.063	0.731	0.115	0.812	0.522	0.542	0.725	0.918	0.810
Overall	0.992	0.919	0.954	0.614	0.888	0.726	0.861	0.924	0.890	0.747	0.900	0.817

Table 4. *System scores for Co-reference Resolution on i2b2_Partners corpus*

Coref type	BCUBED			CEAF		
	P	R	F	P	R	F
Coref test	0.98	0.88	0.927	0.779	0.899	0.835
Coref person	0.88	0.66	0.754	0.24	0.545	0.333
Coref problem	0.9	0.88	0.89	0.706	0.748	0.726
Coref treatment	0.91	0.88	0.895	0.786	0.779	0.782
Overall	0.92	0.86	0.889	0.582	0.746	0.654

Table 5. *End-to-end Co-reference evaluations on ODIE_CLINICAL corpus*

Coref type	BCUBED			CEAF		
	P	R	F	P	R	F
Coref test	0.92	0.97	0.944	0.852	0.767	0.807
Coref problem	0.58	0.91	0.708	0.575	0.424	0.488
Coref treatment	0.95	0.95	0.95	0.95	0.925	0.938
Overall	0.75	0.92	0.826	0.751	0.53	0.622

Table 6. *End-to-end Co-reference evaluations on ODIE_PATH corpus*

6. DISCUSSION AND CONCLUSION

In the Concept Annotation experiment, the micro averaged F-scores are about 85%. Tests are the most difficult entities to recognize, due to much lower frequencies or larger sizes than other entity types, and abundant variants. In the End-to-end Co-reference evaluations, the micro averaged F-scores decline to less than 66%, in the ODIE corpora, using the CEAF evaluation metric, but remain about 88.9% and 82.6% in BCUBED evaluation metric. The probable reasons are the errors in the concept annotation, the limited number of instances, and the poorer grammatical structure in the sentences compared to the i2b2 corpora.

The results of Co-reference Resolution on the i2b2 corpora show over 94% on the micro averaged F-scores in BCUBED evaluation metric and over 81% in CEAF evaluation metric. This suggests that such a system can be used for co-reference resolution in clinical texts and achieve reliable performance on a large corpus.

In future work, for a small corpus, other classification strategies (e.g., rule-based approaches) can be utilized to improve the performance of the system.