Appendix II. Corpora annotation details

The coreference annotation task involves linking pairs of concept mentions that relate through an identity relationship. Coreference chains are generated by applying the transitivity rule to linked pairs.

Annotation Tasks for the 2011 i2b2/VA Challenge

The 2011 i2b2/VA challenge built on the problem, test, and treatment concept mentions annotated in 2010 i2b2/VA challenge and augmented these categories with person and pronoun categories. Person category included proper names, personal pronouns, and names of groups of people. The pronoun category included non-person pronouns.

Annotation Guidelines, Schema and Tool

i2b2/VA coreference annotation guidelines were based on the ODIE guidelines.[1] The i2b2/VA corpus was annotated over a seven month period, through the combined efforts of 23 annotators. Eleven of these annotators were clinicians; 12 were non-clinicians. All annotators were trained for 1.5-2.5 hours on the annotation guidelines and the annotation tool before starting annotation.

Annotator Agreement Metrics

We measured the inter-annotator agreement (IAA) on coreference pairs, given concept mentions, using the formula:[2]

$$IAA = \frac{2 * Matches}{(2 * Matches + Non - Matches)}$$

where *Matches* correspond to agreements and *Non-Matches* correspond to disagreements.

Evaluation metrics were computed by constructing a 2x2 table for each individual category. Overall IAA are micro-averaged across all classes for concept level IAA, and across all relations for coreference pair IAA (see Table 3).

$$IAA\_Micro\_Average = \frac{\sum_{i=1}^{M} 2 * Matches_i}{\sum_{i=1}^{M} (2 * Matches_i + Non - Matches_i)}$$

where *M* is the total number of classes.

We limited IAA analysis to pairs, in order to measure the reliability of the manual linking task; chains were automatically generated from pairs and were therefore excluded from IAA analysis. These metrics differ from those used to assess system performance of 2011 i2b2/VA challenge submissions in three important ways. First, these agreement metrics represent the independent human annotation task on raw documents where ground truth concepts for problems, treatments, and tests were provided to annotators. Annotators identified and marked mentions of persons and were allowed to modify pre-annotated pronoun annotations. Annotators were also encouraged to report on incorrect problem, test, treatment annotations for potential curation. Second, annotators followed a procedure that involved a minimal amount of effort to link anaphor-antecedent pairs and to create coreference chains. Third, these metrics differ from system evaluation metrics since they are geared towards evaluating manual annotation tasks.

Common Sources of Annotators Disagreement

From a qualitative perspective there were four common sources of disagreement between annotators for the 2011 coreference annotation task. These included: 1) problems related to use of the annotation tool and the visualization of overlapping mentions; 2) overlinking; 3) underlinking; and 4) chaining of unrelated entities. In the first case, reuse of the 2010 i2b2 corpus that contained overlapping mention annotations that included articles and pronouns were difficult to visualize using the annotation tool. For example, where the entire noun phrase “her fever” is marked as a problem, the pronoun “her” may be missed by the annotator. Overlinking of mentions often occurred in situations where a temporal relationship between distinct events existed. For example, “temp” and “temperature” in “Temp on admit was 102” and “her temperature today had lowered to 99” are separate events and are not coreferential. However, the definition of a distinct event was subjective and often left up to the annotator to determine. Our guidelines specifically defined excluding set-subset or part-whole coreferential relationships, but there was often gray area that was subjective despite our attempts to clearly define and explicate inclusions and exclusions. Underlinking also occurred in situations where annotators lacked sufficient medical knowledge or a coreferential relationship was not obvious due to nuances of the medical sublanguage used to describe the same entity or in situations where an acronym could not be disambiguated. This type of situation occurred frequently for problems, treatments, and tests. It is not uncommon, for example, for a provider to document the presence of a medical device, medication, a test, or problem using some generic form, or use an acronym or short form that refers to the same entity. For example, consider the sentence “patient reports with subgleal bleed and a subdural bleed”. Later in the same document there is mention of [the bleed] causing “midline shift”, which may only be a result of subdural bleeding. Surprisingly, both clinician and non-clinician annotators struggled with this particular issue and a certain degree of subjectivity was required in these kinds of situations. Underlinking was also more common in longer documents where it was more difficult for the annotators to recall information and make coreferential links across an entire document.

One final problem that occurred less often included inappropriate chaining of unrelated mentions. In these cases annotators could inadvertently include unrelated mentions in the same chain. For example, annotators linked [liver cancer] to [cancer]; they also linked [breast cancer] to [cancer]. Since we used automatic methods for creating chains out of mention pairs, this could result in [liver cancer] and [breast cancer] inadvertently ending up in the same chain.

Appendix III. Evaluation metrics for mention extraction

Following the evaluation methodology of the fourth i2b2 challenge,[3] we evaluated mention extraction using the precision (P), recall (R), and F-measure (F) metrics. These metrics are computed based on the true positives (TPs), false positives (FPs), and false negatives (FNs) retrieved by a system. We defined the TP, FP, and FN differently for mentions that exactly overlapped the gold standard mentions (i.e., exact overlap), and for mentions that at least partially overlapped the gold standard mentions (i.e., at least partial overlap).

For exact overlap, we defined TP, FP, and FN as:

- TP: system mention annotation exactly agreed with gold standard mention annotation, in both token offset and semantic category.
- FP: system mention annotation did not exactly agree with gold standard annotation, in either token offset or semantic category.
- FN: gold standard annotation did not exactly agree with system mention annotation, in either token offset or semantic category.

For at least partial overlap we defined TP, FP, and FN as:

- TP: system mention annotations for token offset partially overlapped the gold standard mention annotation. The system mention annotation for semantic category exactly agreed with the gold standard mention annotation.
- FP: system mention annotation for token offset did not overlap with the gold standard annotation, or the system mention annotation for semantic category did not exactly agree with the gold standard mention annotation.
- FN: gold standard mention annotation for token offset did not overlap with the system annotation, or the gold standard mention annotation for semantic category did not exactly agree with the system mention annotation.

Uzuner et al., Evaluating the state of the art in coreference resolution for electronic medical records

References:

1. Savova G, Chapman WW, Zheng J, *et al.* Anaphoric relations in the clinical narrative: corpus creation. Journal of American Medical Informatics Association 2011;18:459-65.
2. Roberts A, Gaizauskas R, Hepple M, *et al.* Building a semantically annotated corpus of clinical texts. Journal of Biomedical Informatics 2009;42:950-66.
3. Uzuner O, South B, Shen S, *et al.* 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical texts. Journal of American Medical Informatics Association 2011;18:552-56.

Table 1: 2011 i2b2/VA challenge participants

Number	Author name	Task	Participating Organization	Country
1	Anick et al.	Task 1C	Brandeis University	U.S.
2	Benajiba et al.	Task 1B, 1C	Philips Research North America	U.S.
3	Cai et al.	Task 1A, 1B, 1C	Heidelberg Institute for Theoretical Studies gGmbH	Germany
4	Dai et al.	Task 1C	Academia Sinica Yuan Ze University National Tsing Hua University	Taiwan
5	Glinos	Task 1B, 1C	Advanced Text Analytics, LLC	U.S.
6	Gooch	Task 1B, 1C	Centre for Health Informatics, City University	England
7	Grouin et al.	Task 1A, 1B, 1C	LIMSI-CNRS Universite Paris-Sud	France
8	Guillen	Task 1C	California State University San Marcos	U.S.
9	Hinote et al.	Task 1B, 1C	University of Houston - Downtown	U.S.
10	Jindal et al.	Task 1C	University of Illinois at Urbana-Champagne	U.S.
11	Jonnalagadda et al.	Task 1C	Mayo Clinic Georgetown University	U.S.
12	Lan et al.	Task 1A, 1B, 1C	East China Normal University	China
13	Patrick et al.	Task 1C	University of Sydney	Australia
14	Rink et al.	Task 1B, 1C	University of Texas Dallas	U.S.
15	Wang et al.	Task 1C	Arizona State University	U.S.
16	Ware et al.	Task 1C	West Virginia University Medquist	U.S.
17	Weissenbacher et al.	Task 1C	Toyota Technology Institute, Japan	Japan
18	Xu et al.	Task 1C	Microsoft Research Asia Beihang University Tsinghua Univesrity Shanghai Jiaotong University	China
19	Yan et al.	Task 1C	Harbin Institute of Technology	China
20	Yang et al.	Task 1C	Open University Lero, University of Limerick	England Ireland

Table 2: Inter-annotator agreement results

IAA Mention Extraction												
	BIDMC		**PH**		**UPMC Discharge**		**UPMC Progress**		**UPMC Discharge & Progress**		**Total**	
	Exact	At least partial	Exact	At least partial	Exact	At least partial	Exact	At least partial	Exact	At least partial	Exact	At least partial
Problem	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1	1
Person	0.85	0.90	0.95	0.97	0.89	0.94	0.84	0.91	0.87	0.93	0.9	0.94
Pronoun	0.96	0.96	0.97	0.97	0.96	0.96	0.95	0.95	0.96	0.96	0.96	0.96
Test	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1	1
Treatment	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1	1
Overall	**0.98**	**0.98**	**0.98**	**0.99**	**0.96**	**0.97**	**0.95**	**0.97**	**0.96**	**0.97**	**0.97**	**0.98**

IAA Coreference Resolution									
	BIDMC		**PH**		**UPMC Discharge**		**UPMC Progress**		**UPM**
	Exact	At least partial	Exact	At least partial	Exact	At least partial	Exact	At least partial	Exac
Problem	0.65	0.65	0.64	0.64	0.72	0.72	0.59	0.59	0.67
Person	0.79	0.84	0.91	0.92	0.84	0.87	0.72	0.80	0.79
Pronoun	0.65	0.66	0.44	0.44	0.59	0.60	0.51	0.53	0.57
Test	0.36	0.36	0.44	0.45	0.38	0.39	0.41	0.41	0.40
Treatment	0.49	0.49	0.54	0.54	0.58	0.58	0.68	0.68	0.62
Overall	**0.66**	**0.68**	**0.80**	**0.81**	**0.77**	**0.79**	**0.67**	**0.73**	**0.73**

Table 3: Task 1A mention extraction evaluation using precision, recall, and F-measure on at least partial and exact mention overlap

	At least partial overlap			**Exact overlap**		
	P	R	F	P	R	F
Lan et al.	0.832	0.661	0.737	0.728	0.579	0.645
Grouin et al.	0.692	0.789	0.737	0.584	0.667	0.623
Cai et al.	0.518	0.571	0.544	0.450	0.496	0.472

Table 4: Statistical significance results for teams participating in Task 1A. Only the upper diagonal is marked with the p-value results.

	At least partial overlap		**Exact overlap**	
	Lan et al.	Cai et al.	Lan et al.	Cai et al.
Grouin et al.	0.189	0.01	0.267	0.01
Lan et al.	■	0.01	■	0.01

Table 5: Statistical significance results for teams participating in Task 1B. Only the upper diagonal is marked with the p-value results.

	Rink et al.	Cai et al.	Grouin et al.	Hinote et al.	Lan et al.	Gooch	Benajiba et al.
Glinos	1	1	0.01	0.01	0.01	0.01	0.01
Rink et al.		1	0.01	0.01	0.01	0.01	0.01
Cai et al.			0.01	0.109	0.01	0.01	0.01
Grouin et al.				0.02	1	0.01	0.03
Hinote et al.					0.089	0.505	0.99
Lan et al.						0.01	1
Gooch							1