Semantic Annotation of Clinical Text: The CLEF Corpus

Angus Roberts, Robert Gaizauskas, Mark Hepple, George Demetriou, Yikun Guo, Andrea Setzer, Ian Roberts

Department of Computer Science, University of Sheffield
Regent Court, 211 Portobello, Sheffield, UK S1 4DP
E-mail: initial.surname@dcs.shef.ac.uk

Abstract

A significant amount of important information in Electronic Health Records (EHRs) is often found only in the unstructured part of patient narratives, making it difficult to process and utilize for tasks such as evidence-based health care or clinical research. In this paper we describe the work carried out in the CLEF project for the semantic annotation of a corpus to assist in the development and evaluation of an Information Extraction (IE) system as part of a larger framework for the capture, integration and presentation of clinical information. The CLEF corpus consists of both structured records and free text documents from the Royal Marsden Hospital pertaining to deceased cancer patients. The free text documents are of three types: clinical narratives, radiology reports and histopathology reports. A subset of the corpus has been selected for semantic annotation and two annotation schemes have been created and used to annotate: (i) a set of clinical entities and the relations between them, and (ii) a set of annotations for time expressions and their temporal relations with the clinical entities in the text. The paper describes the make-up of the annotated corpus, the semantic annotation schemes used to annotate it, details of the annotation process and of inter-annotator agreement studies, and how the annotated corpus is being used for developing supervised machine learning models for IE tasks.

1. Introduction

Although large parts of the patient electronic health care record exist as structured data, a significant proportion exists as unstructured free texts. This is not just the case for legacy records. Much of pathology and imaging reporting is recorded as free text, and a major component of any UK medical record consists of letters written from the secondary to the primary care physician (GP). These documents contain information of value for day-to-day patient care and of potential use in research. For example, narratives record why drugs were given, why they were stopped, the results of physical examination, and problems that were considered important when discussing patient care, but not important when coding the record for audit. Clinical researchers could be assisted in hypothesis formation (for subsequent verification in clinical trials) if they could get answers aggregated across all NHS patient records to questions such as:

> *How many patients with stage 2 adenocarcinoma who were treated with tamoxifen were symptom-free after 5 years?*

Doctors could also benefit for treating individual patients if they could get concise summaries of patients' clinical histories or if they had access to histories of similar patients elsewhere.

CLEF (Rector et al. 2003) uses IE technology to make information available for integration with the structured record, and thus to make it available for clinical care and research (Harkema et al. 2005). Although some IE research has focused on unsupervised methods of developing systems, as in the earlier work of Riloff (1996), most practical IE still needs data that has been manually annotated with events, entities and relationships. This data serves three purposes. Firstly, an analysis of human annotated data focuses and clarifies requirements. Secondly, it provides a gold standard against which to assess results. Thirdly, it provides data for system development: extraction rules may be created either automatically or by hand, and statistical models of the text may be built by machine learning algorithms.

This paper reports on the construction of a gold standard corpus for the CLEF project, in which clinical documents are annotated both with multiple entities and their relationships. To the best of our knowledge, no one has explored the problem of producing a corpus annotated for clinical IE to the depth and to the extent reported here. Our annotation exercise uses a large corpus, covers multiple text types, and involves over 20 annotators. We examine two issues of pertinence to the annotation of clinical documents: the use of domain knowledge; and the applicability of annotation to different sub-genres of text. Results are encouraging, and suggest that a rich corpus to support IE in the medical domain can be created. An earlier description of the CLEF corpus was reported in (Roberts et al. 2007). The current paper provides more details, including details of the temporal annotation (not reported at all earlier), figures on the distribution of entity and relation types across the corpus, and inter annotator agreement scores for the completed corpus.

The next section of this paper summarises the literature about annotated biomedical corpora. The following section describes the design of the CLEF corpus, describing the selection of documents for gold standard semantic annotation and the entities and relationships with which the gold standard is annotated. Next the annotation methodology is described, including a discussion of the development of annotation guidelines and an assessment of the consistency of human annotations. The following sections present inter annotator agreement scores for the finished corpus, and figures on the distribution of entity and relation types by document type across the corpus. Finally we mention on-going use of the corpus for training and evaluation of our supervised machine learning IE system.

2. Annotated Corpora for Biomedical Research

Semantically annotated corpora are becoming increasingly common within biomedical information extraction research, with annotation levels gradually expanding over the years. For example, the GENIA corpus of Medline abstracts has been annotated with information about biological entities (Kim et al. 2003) with annotations about biological events added to (part of) it at a later stage (Kim et al. 2008). Other semantically annotated corpora developed for the purpose of providing training and evaluation material for IE systems include:

- The PennBioIE corpus of ~2300 Medline abstracts, in the domains of molecular genetics of oncology and inhibition of enzymes of the CYP450 class annotated for biomedical entity types and parts-of-speech, some of which have also been annotated for Penn Treebank style syntactic structure (Mandel, 2006);
- The Yapex corpus of 200 Medline abstracts annotated for protein names (Franzén et al. 2002);
- Those developed within the BioText project for disease-treatment relation classification (Rosario and Hearst, 2004) and protein-protein interaction classification (Rosario and Hearst, 2005).

In addition corpora have been available in order to provide data sets for research competitions such as:

- Biocreative (the GENETAG corpus containing 15,000 sentences with gene/protein names annotated – Tanabe et al 2005)
- the TREC Genomics Track, which ran from 2003-2007 and for which a variety of datasets and tasks were developed (http://ir.ohsu.edu/genomics/).
- the LLL05 challenge task, which supplied training and test data for the task of identifying protein/gene interactions in sentences from Medline abstracts (Nédellec, 2005).

All of the above corpora consist of texts drawn from the research literature, in most cases from the biology research literature. This is due at least in part to the difficulty of getting access to clinical text for research purposes. To our knowledge the only other work in the area of corpus annotation for clinical information retrieval and extraction is:

- The corpus prepared and released for the Computational Medicine Challenge (Pestian et al 2007). This corpus consists of 1954 (978 training, 976 test) radiology reports annotated with ICD-9-CM codes, the challenge being the text classification challenge of automatically coding the unseen test data.
- The ImageCLEFmed 2005 and 2006 image test collections which consist of ~50,000 images with associated textual annotations (case descriptions, imaging reports) and in some cases metadata (e.g. DICOM labels), together with query topics and relevance judgements (Hersh et al 2006; Müller et all 2007). While intended to support medical image retrieval research, the textual component of this resource could have purely language processing applications.
- Ogren et al.'s (2006) work on annotating disorders within clinic notes; and
- The I2B2 challenges, which have so far provided training and evaluation data for de-identification of discharge summaries and for the identification of smoking status from discharge summaries (challenge 1); and for identification of obesity and co-morbidities from discharge summaries annotated at the document level (https://www.i2b2.org/NLP/).

What differentiates CLEF from the annotation exercises mentioned above is that (1) it is the only corpus annotated with information about clinical entities and their relations as well as with temporal information about the clinical entities and time expressions occurred in patient narratives and (2) it is the only corpus to contain clinic notes, radiology reports and histopathology reports together with associated structured data.

3. Design of the CLEF Corpus

Our development corpus comes from CLEF's main clinical partner, the Royal Marsden Hospital, a large specialist oncology centre. The entire corpus consists of both the structured records and free text documents from 20234 deceased patients. The free text documents consist of three types: clinical narratives (with sub-types as shown in Table 1); histopathology reports; and imaging reports. Patient confidentiality is ensured through a variety of technical and organisational measures, including automatic pseudonymisation and manual inspection. Approval to use this corpus for research purposes within CLEF was sought and obtained from the Thames Valley Multi-centre Research Ethics Committee (MREC).

3.1 Gold Standard Document Sampling

Given the expense of human annotation, the gold standard portion of the corpus has to be a relatively small subset of the whole corpus of 565000 documents. In order to avoid events that are either rare or outside of the main project requirements, it is restricted by diagnosis, and only considers documents from those patients with a primary diagnosis code in one of the top level sub-categories of ICD-10 Chapter II (neoplasms). In addition, it only contains those sub-categories that cover more than 5% of narratives and reports. The gold standard corpus consists of two portions, selected for slightly different purposes.

3.1.1 Whole patient records

Two applications in CLEF involve aggregating data across a single patient record. The CLEF chronicle builds a chronological model for a patient, integrating events from both the structured and unstructured record (Rogers et al 2006). CLEF report generation creates aggregated and natural language reports from the chronicle (Hallet et al 2006). These two applications require whole patient records for development and testing. Two whole patient records were selected for this portion of the corpus, from two of the major diagnostic categories, to give median numbers of documents, and a mix of document types and lengths. Each record consists of nine narratives, one radiology report and seven histopathology reports, plus associated structured data.

3.1.2 Stratified random sample

The major portion of the gold standard serves as development and evaluation material for IE. In order to ensure even training and fair evaluation across the entire corpus, the sampling of this portion is randomised and stratified, so that it reflects the population distribution along various axes. Table 1 shows the proportions of clinical narratives along two of these axes. The random sample consists of 50 each of clinical narratives, histopathology reports, and imaging reports.

Narrative subtype	% of standard
To GP	49
Discharge	17
Case note	15
Other letter	7
To consultant	6
To referrer	4
To patient	3

Neoplasm	% of standard
Digestive	26
Breast	23
Haematopoetic	18
Respiratory etc	12
Female genital	12
Male genital	8

Table 1: % of narratives in random sample

3.2 Annotation Schema: Clinical Information

The CLEF gold standard is a semantically annotated corpus. We are interested in extracting the main semantic entities from text. By *entity*, we mean some real-world concept referred to in the text such as the drugs that are mentioned, the tests that were carried out etc. We are also interested in extracting the *relationships* between entities: the condition indicated by a drug, the result of an investigation etc.

Annotation is anchored in the text. Annotators mark spans of text with a type: drug, locus and so on. Annotators may also mark words that modify spans (such as negation), and mark relationships as links between spans. Two or more spans may refer to the same thing in the real world, in which case they *co-refer*. Co-referring CLEF entities are linked by the annotators.

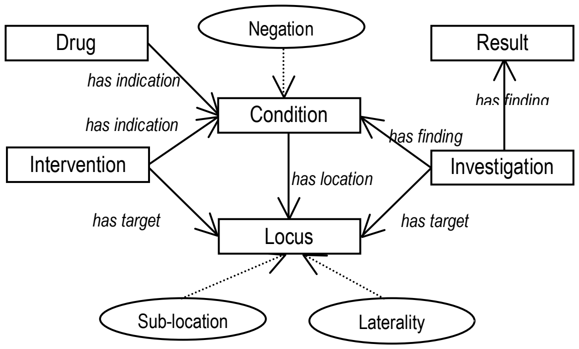

Figure 1: CLEF annotation schema. Rectangles: entities; ovals: modifiers; solid lines: relationships.

The types of annotation are described in a schema, shown in Figure 1. The CLEF entities and relations are also listed in Tables 2 and 3, along with descriptions and examples.

The schema has been based on a set of requirements developed between clinicians and computational linguists in CLEF. The schema types are mapped to types in the UMLS semantic network, which enables us to utilize UMLS vocabularies in entity recognition. For the purposes of annotation, the schema is modeled as a Protégé-Frames ontology (Gennari et al. 2003). Annotation is carried out using an adapted version of the Knowtator plugin for Protégé (Ogren 2006). This was chosen for its handling of relationships, after evaluating several such tools.

3.3 Annotation Schema: Temporal Information

Information from structured data and clinical narratives is integrated to build a *patient chronicle*, i.e. a coherent overview of the significant events in the patients' medical history, such as their condition, diagnosis and treatment over the period of care. This process involves extracting temporal information about events from the narratives, and using this and other information to map the events extracted from the narratives onto their corresponding, time-stamped, events in the structured data wherever possible. The aim of the gold standard is to provide the temporal links (called CTlinks for CLEF Temporal link) between TLCs (Temporally Located CLEF entities, which comprise CLEF investigations, interventions and conditions) and temporal expressions. Temporal expressions include dates and times (both absolute and relative), as well as durations, as specified in the TimeML (2004) TIMEX3 standard. CTlinks types include, for example, before, after, overlap, includes (for a full list see Table 9). Our scheme requires annotation of only those temporal relations holding between TLCs and the date of the letter (Task A), and between TLCs and temporal expressions appearing in the same sentence (Task B). These tasks are similar to, but not identical with, those addressed by the TempEval challenge within SemEval 2007 (Verhagen et al. 2007).

4. Annotation Methodology

The annotation methodology follows established natural language processing standards (Boisen et al. 2000). Annotators work to agreed guidelines; documents are annotated by at least two annotators; documents are only used where agreement passes a threshold; differences are resolved by a third experienced annotator. These points are discussed further below.

Entity type	Description	Example
Condition	Symptom, diagnosis, complication, conditions, problems, functions and processes, injury	*This patient has had a lymph node biopsy which shows melanoma in his right groin. It is clearly secondaries from the melanoma on his right second toe.*
Intervention	Action performed by doctor or other clinician targeted at a patient, **Locus**, or **Condition** with the objective of changing (the properties) of, or treating, a **Condition**.	*Although his PET scan is normal he does need a groin dissection* *We agreed to treat with DTIC, and then consider radiotherapy*.
Investigation	Interaction between doctor and patient or **Locus** aimed at measuring or studying, but not changing, some aspect of a **Condition**. **Investigations** have findings or interpretations, whereas **Interventions** usually do not.	*This patient has had a lymph node biopsy … Although his PET scan is normal he does need a groin dissection. We will perform a CT scan to look at the left pelvic side wall …*
Result	The numeric or qualitative finding of an **Investigation**, excluding **Condition**	*Although his PET scan is normal…* Other examples include the numeric values of tests, such as "80mg".
Drug or device	Usually a drug. Occasionally, medical devices such as suture material and drains will also be mentioned in texts.	*This (pain) was initially relieved by co-codamol*
Locus	Anatomical structure or location, body substance, or physiologic function, typically the locus of a **Condition**.	*This patient has had a lymph node biopsy which shows melanoma in his right groin … It is clearly secondaries from the melanoma on his right second toe. Although his PET scan is normal he does need a groin dissection. We will perform a CT scan to look at the left pelvic side wall*

Table 2: CLEF Entities

Relation type	1st arg type	2nd arg type	Description	Example
has_target	**Investigation, Intervention**	**Locus**	Relates an intervention or an investigation to the bodily locus at which it is targetted.	*This patient has had a [arg2 lymph node] [arg1 biopsy]* *… he does need a [arg2 groin] [arg1 dissection]*
has_finding	**Investigation**	**Condition, Result**	Relates a condition to an investigation that demonstrated its presence, or a result to the investigation that produced that result.	*This patient has had a lymph node [arg1 biopsy] which shows [arg2 melanoma]* *Although his [arg1 PET] scan is [arg2 normal]*
has_indication	**Drug or device, Intervention, Investigation**	**Condition**	Relates a condition to a drug, intervention, or investigation that is targetted at that condition	*Her facial [arg2 pain] was initially relieved by [arg1 co-codamol]*
has_location	**Condition**	**Locus**	Relationship between a condition and a locus: describes the bodily location of a specific condition. May also describe the location of malignant disease in lymph nodes, relating an involvement to a locus.	*… a biopsy which shows [arg1 melanoma] in his right [arg2 groin]* *It is clearly secondaries from the [arg1 melanoma] on his right [arg2 second toe]* *Her[arg2 facial] [arg1 pain] was initially relieved by co-codamol*
Modifies	**Negation signal**	**Condition**	Relates a condition to its negation or uncertainty about it	*There was [arg1 no evidence] of extra pelvic [arg2 secondaries]*
Modifies	**Laterality signal**	**Locus, Intervention**	Relates a bodily locus or intervention to its sidedness: *right*, *left*, *bilateral*.	*… on his [arg1 right] [arg2 second toe]* [arg1 *right*] [arg2 *thoracotomy*]
Modifies	**Sub-location signal**	**Locus**	Relates a bodily locus to other information about the location: *upper*, *lower*, *extra*, etc.	*[arg1 extra] [arg2 pelvic]*

Table 3: CLEF Relations

4.1 Annotation Guidelines

Consistency is critical to the quality of a gold standard. It is important that all documents are annotated to the same standard. Questions regularly arise when annotating. For example, should multi-word expressions be split? Should “myocardial infarction” be annotated as a condition, or as a condition and a locus? To ensure consistency, a set of guidelines is provided to annotators. These describe in detail what should and should not be annotated; how to decide if two entities are related; how to deal with co-reference; and a number of special cases. The guidelines also provide a sequence of steps, a recipe, which annotators should follow when working on a document. This recipe is designed to minimise errors of omission. The guidelines themselves were developed through a rigorous, iterative process, which is described below.

4.2 Double Annotation

A singly annotated document can reflect many problems: the idiosyncrasies of an individual annotator; one-off errors made by a single annotator; annotators who consistently under-perform. There are many alternative annotation schemes designed to overcome this, all of which involve more annotator time. Double annotation is a widely used alternative, in which each document is independently annotated by two annotators, and the sets of annotations compared for agreement.

4.3 Agreement Metrics

We measure agreement between double annotated documents using *inter annotator agreement* (IAA, shown below).

IAA = matches / (matches + non-matches)

We report IAA as a percentage. Overall figures are macro-averaged across all entity or relationship types. Entity IAA may be either “relaxed” or “strict”. In relaxed IAA, partial matches, i.e. overlaps, are counted as a half match. In strict IAA, partial matches do not count to the score. Together, these show how much disagreement is down to annotators finding similar entities, but differing in the exact spans of text marked. We used both scores in development, but provide a set of final strict IAAs for the corpus. Results given below explicitly state the score being used.

Two variations of relationship IAA were used. First, all relationships found were scored. This has the drawback that an annotator who failed to find a relationship because they had not found one or both the entities would be penalized. To overcome this, a Corrected IAA (referred to as CIAA) was calculated, including only those relationships where both annotators had found the two entities involved. This allows us to isolate, to some extent, relationship scoring from entity scoring.

4.4 Difference Resolution

Double annotation can be used to improve the quality of annotation, and therefore the quality of statistical models trained on those annotations. This is achieved by combining double annotations to give a set closer to the "truth" (although it is generally accepted as impossible to define an "absolute truth" gold standard in an annotation task with the complexity of CLEF's). The resolution process is carried out by a third experienced annotator. All agreements from the original annotators are accepted into a consensus set, and the third annotator adjudicates on differences, according to a set of strict guidelines. In this way, annotations remain at least double annotated.

4.5 Developing the Guidelines

The guidelines were developed and refined using an iterative process, designed to ensure their consistency. This is shown in Figure 2. Two qualified clinicians annotated different sets of documents in 5 iterations (covering 31 documents in total). The relaxed IAA and CIAA for these iterations are shown in Table 4. As can be seen, entity relaxed IAA remains consistently high after the 5 iterations, after which very few amendments were required on the guidelines. Relation CIAA does not appear so stable on iteration 5. Difference analysis showed this to be due to a single, simple type of disagreement across a limited number of sentences in one document. Scoring without this document gave a 73% CIAA.

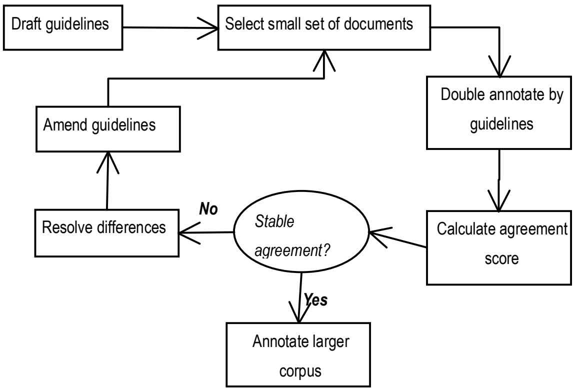

Figure 2: Iterative development of guidelines

		Debug iteration				
		1	2	3	4	5
Entities	Matches	244	244	308	462	276
	Partial match	2	6	22	6	1
	Non-matches	45	32	93	51	22
	Relaxed IAA	**84**	**87**	**74**	**89**	**92**
Relationships	Matches	170	78	116	412	170
	Partial match	3	5	14	6	1
	Non-matches	31	60	89	131	103
	Corrected IAA	**84**	**56**	**56**	**75**	**62**

Table 4: Relaxed IAA and CIAA (%) for each development iteration.

4.6 Annotator Expertise

In order to examine how easily the guidelines could be applied by other annotators with varying levels of expertise, we also gave a batch of documents to our development annotators, another clinician, a biologist with some linguistics background, and a computational linguist. Each was given very limited training. The resultant annotations were compared with each other, and with a consensus set created from the two development annotators. The relaxed IAA matrix for this group is shown in Table 5. This small experiment shows that even with very limited training, agreement scores that approach acceptability are achievable. A difference analysis suggested that the computational linguist was finding more pronominal co-references and verbally signaled relations than the clinicians, but that unsurprisingly, the clinicians found more relations requiring domain knowledge to resolve. A combination of both linguistic and medical knowledge appears to be best.

This difference reflects a major issue in the development of the guidelines: the extent to which annotators should apply domain specific knowledge to their analysis. Much of clinical text can be understood, even if laboriously and simplistically, by a non-clinician armed with a medical dictionary. The basic meaning is exposed by the linguistic constructs of the text. Some relationships between entities in the text, however, require deeper understanding. For example, the condition for which a particular drug was given may be unclear to the non-clinician. In writing the guidelines, we decided that such relationships should be annotated, although this requirement is not easy to formulate as specific rules.

D2	77				
C	67	68			
B	76	80	69		
L	67	73	60	69	
Consensus	**85**	**89**	**68**	**78**	**73**
	D1	D2	C	B	L

Table 5: Relaxed IAA (%) for entities. D1 and D2: development annotators; C: clinician; B: biologist with linguistics background; L: computational linguist

4.7 Annotation: Training and Consistency

In total, around 25 annotators were involved in guideline development and annotation. They included practicing clinicians, medical informaticians, and final year medical students. They were each given an initial 2.5 hours training session.

After the initial training session, annotators were given two training batches to annotate, which comprised documents originally used in the debugging exercise, and for which consensus annotations had been created. Relaxed IAA and CIAA were computed between annotators, and against the consensus set. These figures allowed us to identify and offer remedial training to under-performing annotators and to refine the guidelines further.

4.8 Annotation of Temporal Information

This work is still at a preliminary stage. To date ten patient letters (narrative data) for a number of patients have been annotated in accordance with the scheme described in Section 3.3 above, which we still view as under development. A second annotator is currently re-annotating as part of the guideline development phase (see Figure 2). Temporal annotation is done through a combination of manual and automatic methods. TLCs were imported from the part of the corpus already annotated with clinical entities. Temporal expressions were annotated and normalized to ISO dates by the GUTime tagger (Mani and Wilson 2000), developed at Georgetown University, which annotates in accordance with the TIMEX3 standard and also recognizes a variety of temporal modifiers and European date formats. After these automatic steps, we manually annotate the temporal relations holding between TLCs and the date of the letter (Task A), and between TLCs and temporal expressions appearing in the same sentence (Task B**).**

5. Inter annotator agreement

We have calculated IAA for the double annotations across the complete stratified random portions of the gold standard, for each document type. Table 6 shows the strict IAA for entities, and Table 7 shows both the IAA and CIAA for relationships.

Entity	**Narratives**	**Histopath.**	**Radiology**
Condition	81	67	77
Drug or device	84	59	32
Intervention	64	57	43
Investigation	77	56	70
Locus	78	71	75
Result	69	29	48
Laterality	95	88	91
Negation	67	71	65
Sub-location	63	29	36
Overall	**77**	**62**	**69**

Table 6: Strict IAA (%) for entities across the stratified random corpus

Note that the final gold standard consists of a consensus of the double annotation, created by a third annotator. Systems trained and evaluated with the gold standard use this consensus. The IAAs given do not therefore provide an upper bound on system performance, but an indication of how hard a recognition task is. Table 7 illustrates that relation annotation is highly dependent on entity annotation: CIAA, corrected for entity recognition, is significantly higher than uncorrected IAA.

	Narratives		**Histopath.**		**Radiology**	
Relation	**IAA**	**CIAA**	**IAA**	**CIAA**	**IAA**	**CIAA**
has_finding	48	76	26	69	33	55
has_indication	35	51	15	30	14	22
has_location	59	80	44	70	45	77
has_target	45	64	20	47	67	81
laterality_mod	70	93	70	89	55	80
negation_mod	63	90	67	100	51	94
sub_loc_mod	52	98	29	100	32	93
Overall	**52**	**75**	**36**	**72**	**43**	**76**

Table 7: IAA and corrected IAA (%) for relationships across the stratified random corpus

6. Distribution of semantic annotations

The distribution of annotations for CLEF entities and relations in the stratified random portion of the corpus (50 documents of each type) is shown in Table 8.

CLEF stratified random corpus				
Entity	Narra-tives	Histopath-ology	Radiol-ogy	Total
Condition	429	357	270	1056
Drug or device	172	12	13	197
Intervention	191	53	10	254
Investigation	220	145	66	431
Laterality	76	14	85	175
Locus	284	357	373	1014
Negation	55	50	53	158
Result	125	96	71	292
Sub-location	49	77	125	251
Relation				
has_finding	233	263	156	652
has_indication	168	47	12	227
has_location	205	270	268	743
has_target	95	86	51	232
laterality_mod	73	14	82	169
negation_mod	67	54	59	180
sub_loc_mod	43	79	125	247

Table 8: Distribution of annotations by document type for entities and relations (clinical IE).

The distribution of annotations for the different subtypes of CTLinks, TLCs and time expressions for the ten development documents annotated so far are shown in Tables 9 and 10. Note that some TLCs are marked as hypothetical. For example in "no palliative chemotherapy or radiotherapy would be appropriate" the terms chemotherapy and radiotherapy are marked as TLCs but clearly have no "occurrence" that can be located in time and hence will not participate in any CTLinks.

CTLink	Task A	Task B
After	5	18
Ended_by	3	0
Begun_by	4	0
Overlap	7	26
Before	5	135
None	4	8
Is_included	31	67
Unknown	6	14
Includes	13	137
Total	78	405

Table 9: Distribution of CTLinks by type for tasks A & B.

TLCs	Not hypothetical	243
	hypothetical	16
	Total	259
Time Expression	Duration	3
	DATE	52
	Total	55

Table 10: Distribution of TLCs and temporal expressions.

7. Using the Corpus

The gold standard corpus is used as input to train an IE system based on SVM classifiers for recognizing both entities and relations. Preliminary results, with models evaluated on a narrative corpus comprising the combined stratified random and whole patient portion, achieve average F-measure 71% for entity extraction over 5 entity types (Roberts et al. 2008). Preliminary results for relation extraction trained with the same corpus, achieve an average F-measure of 70% over 7 relation types, where gold standard entities are provided as input.

8. Conclusion

We have described the CLEF corpus: a semantically annotated corpus designed to support the training and evaluation of information extraction systems developed to extract information of clinical significance from free text clinic notes, radiology reports and histopathology reports. We have described the design of the annotated corpus, including the number of texts it contains, the principles by which they were selected from a large body of unannotated texts and the annotation schema according to which clinical and temporal entities and relations of significance have been annotated in the texts. We also described the annotation process we have undertaken with a view to ensuring, as far as is possible given constraints of time and money, the quality and consistency of the annotation, and we have reported results of inter-annotator agreement, which show that promising levels of inter-annotator agreement can be achieved. We have examined the applicability of annotation guidelines to several clinical text types, and our results suggest that guidelines developed for one type may be fruitfully applied to others. We have also reported the distribution of entity and relation types, both clinical and temporal, across the corpus, giving a sense of how well represented each entity and relation type is in the corpus.

The annotated CLEF corpus is the richest resource of semantically marked up clinical text yet created. Our work has faced several challenges, such as achieving consistent annotation, particularly of relations, across annotators and co-ordinating the work of many annotators at several sites. We do not as yet have persmission to release these materials to the wider language processing community for research purposes. However, we are currently preparating an application requesting this release, to be submitted shortly to the appropriate UK Multi-centre Research Ethics Committee. We are optimistic of success.

9. Acknowledgements

CLEF is funded by the UK Medical Research Council. We would like to thank the Royal Marsden Hospital for providing the corpus; our annotators; and members of CLEF who have helped with clinical expertise and logistics.

10. References

Boisen S, Crystal MR, Schwartz R, Stone R, Weischedel R. (2000). Annotating resources for information extraction. *Proc Language Resources and Evaluation, pp.* 1211--1214.

Franzén K., Eriksson G., Olsson F., Per Lidén L. A. and Cöster J. (2002). Protein names and how to find them. In

International Journal of Medical Informatics special issue on Natural Language Processing in Biomedical Applications.

Gennari J.H., Musen M.A., Fergerson R.W. et al. (2003). The evolution of Protégé: an environment for knowledge-based systems development. *Int J Hum Comput Stud.*, 58, pp. 89--123.

Hallet C., Power R., Scott D. (2006). Summarisation and visualization of e-health data repositories. *Proc 5th UK e-Science All Hands Meeting.*

Harkema H., Roberts I., Gaizauskas R., and Hepple M. (2005). Information extraction from clinical records. *Proc 4th UK e-Science All Hands Meeting.*

Hersh W.R., Müller H., Jensen J., Yang J., Gorman P., Ruch P. (2006). Advancing biomedical image retrieval: development and analysis of a test collection. *J Am Med Inform Assoc.*, 13, pp. 488--496.

Kim J-D, Ohta T, Tateisi Y, Tsujii J. (2003). GENIA corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics,* 19(1), pp. 180--182.

Kim J-D., Tateisi Y, Tsujii J. (2008). Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics,* 9:10.

Mandel M. (2006). Integrated Annotation of Biomedical Text: Creating the PennBioIE corpus. *Text Mining, Ontologies and Natural Language Processing in Biomedicine*, Manchester, UK.

Mani I. and Wilson G. (2000). Processing of News. *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL2000)*, pp. 69--76.

Müller H., Deselaers T., Deserno T. et al (2007). Overview of the ImageCLEFmed 2006 Medical Retrieval and Medical Annotation Tasks. *Evaluation of Multilingual and Multi-modal Information Retrieval.* LNCS 4730/2007, Springer, pp. 595-608.

Nédellec, C. (2005). Learning Language in Logic - Genic Interaction Extraction Challenge. *Proceedings of the ICML05 Workshop on Learning Language in Logic*, Bonn. pp. 31-37.

Ogren P.V., Savova G.K., Buntrock J.D., Chute C.G. (2006). Building and evaluating annotated corpora for medical NLP systems. *Proc AMIA Annual Symposium.*

Ogren, P.V. (2006). Knowtator: a Protégé plug-in for annotated corpus construction. *Proc Human Language Technology*, pp. 273--275.

Pestian JP, Brew C, Matykiewicz PM, Hovermale DJ, Johnson N, Cohen KB, Duch W. (2007). A shared task involving multi-label classification of clinical free text. Proc. ACL BioNLP 2007, Prague.

Rector A., Rogers J., Taweel A. et al. (2003). CLEF: joining up healthcare with clinical and post-genomic research. *Proc 2nd UK e-Science All Hands Meeting.* pp. 264--267.

Riloff E. (1996), Automatically generating extraction patterns from untagged text. *Proc 13th Nat Conf on Artificial Intelligence*, pp. 1044--1049.

Roberts A., Gaizauskas R., Hepple M. et al (2007). The CLEF Corpus: Semantic Annotation of Clinical Text. *AMIA 2007 Proceedings: Biomedical and Health Informatics: From Foundations to Applications to Policy*, pp. 625--629.

Roberts A., Gaizauskas R., Hepple M., Guo Y. (2008). Combining terminology models and statistical methods for entity recognition: an evaluation. *Proc Language Resources and Evaluation. Accepted for publication.*

Rogers J, Puleston C, Rector A. (2006). The CLEF chronicle: patient histories derived from electronic health records. *Proc 22nd Int Conf on Data Engineering Workshops*. p. 109.

Rosario, B. and Hearst, M. (2004). Classifying Semantic Relations in Bioscience Text. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004).

Rosario, B. and Hearst, M. (2005). Multi-way Relation Classification: Application to Protein-Protein Interaction. *HLT-NAACL'05*, Vancouver.

Sager N, Lyman M, Bucknall C, Nhan N, Tick LJ. (1994). Natural language processing and the representation of clinical data. *J Am Med Inform Assoc.* 1(2), pp. 142--160.

Tanabe L., Xie N., Thom L., Matten W. and Wilbur J. (2005). GENETAG: a tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics,* 6, suppl. 1:S3.

TimeML (2004). http://www.cs.brandeis.edu/~jamesp/-arda/ time/.

Verhagen, M., Gaizauskas. R., Schilder, F., Hepple M., Katz, G. and Pustejovsky, J. (2007). SemEval-2007 Task 15: TempEval Temporal Relation Identification. *Proceedings of the 4th International Workshop on Semantic Evaluations*, pp. 75-80.