Extracting Clinical Relationships from Patient Narratives

Angus Roberts, Robert Gaizauskas, Mark Hepple
Department of Computer Science, University of Sheffield,
Regent Court, 211 Portobello, Sheffield S1 4DP
{initial.surname}@dcs.shef.ac.uk

Abstract

The Clinical E-Science Framework (CLEF) project has built a system to extract clinically significant information from the textual component of medical records, for clinical research, evidence-based healthcare and genotype-meets-phenotype informatics. One part of this system is the identification of relationships between clinically important entities in the text. Typical approaches to relationship extraction in this domain have used full parses, domain-specific grammars, and large knowledge bases encoding domain knowledge. In other areas of biomedical NLP, statistical machine learning approaches are now routinely applied to relationship extraction. We report on the novel application of these statistical techniques to clinical relationships.

We describe a supervised machine learning system, trained with a corpus of oncology narratives hand-annotated with clinically important relationships. Various shallow features are extracted from these texts, and used to train statistical classifiers. We compare the suitability of these features for clinical relationship extraction, how extraction varies between inter- and intra-sentential relationships, and examine the amount of training data needed to learn various relationships.

1 Introduction

The application of Natural Language Processing (NLP) is widespread in biomedicine. Typically, it is applied to improve access to the ever-burgeoning research literature. Increasingly, biomedical researchers need to relate this literature to phenotypic data: both to populations, and to individual clinical subjects. The computer applications used in biomedical research, including NLP applications, therefore need to support genotype-meets-phenotype informatics and the move towards translational biology. Such support will undoubtedly include linkage to the information held in individual medical records: both the structured portion, and the unstructured textual portion.

The Clinical E-Science Framework (CLEF) project (Rector et al., 2003) is building a framework for the capture, integration and presentation of this clinical information, for research and evidence-based health care. The project's data resource is a repository of the full clinical records for over 20000 cancer patients from the Royal Marsden Hospital, Europe's largest oncology centre. These records combine structured information, clinical narratives, and free text investigation reports. CLEF uses information extraction (IE) technology to make information from the textual portion of the medical record available for integration with the structured record, and thus available for clinical care and research. The CLEF IE system analyses the textual records to extract entities, events and the relationships between them. These relationships give information that is often not available in the structured record. Why was a drug given? What were the results of a physical examination? What problems were not present? We have previously reported entity extraction in the CLEF IE system (Roberts et al., 2008b). This paper examines relationship extraction.

Extraction of relationships from clinical text is usually carried out as part of a full clinical IE system. Several such systems have been described. They generally use a syntactic parse with domain-specific grammar rules. The Linguistic String project (Sager et al., 1994) used a full syntactic and

BioNLP 2008: Current Trends in Biomedical Natural Language Processing, pages 10–18,
Columbus, Ohio, USA, June 2008. ©2008 Association for Computational Linguistics

clinical sublanguage parse to fill template data structures corresponding to medical statements. These were mapped to a database model incorporating medical facts and the relationships between them. MedLEE (Friedman et al., 1994), and more recently BioMedLEE (Lussier et al., 2006) used a semantic lexicon and grammar of domain-specific semantic patterns. The patterns encode the possible relationships between entities, allowing both entities and the relationships between them to be directly matched in the text. Other systems have incorporated large-scale domain-specific knowledge bases. MEDSYNDIKATE (Hahn et al., 2002) employed a rich discourse model of entities and their relationships, built using a dependency parse of texts and a description logic knowledge base re-engineered from existing terminologies. MENELAS (Zweigenbaum et al., 1995) also used a full parse, a conceptual representation of the text, and a large scale knowledge base.

In other applications of biomedical NLP, a second paradigm has become widespread: the application of statistical machine learning techniques to feature-based models of the text. Such approaches have typically been applied to journal texts. They have been used both for entity recognition and extraction of various relations, such as protein-protein interactions (see, for example, Grover et al (2007)). This follows on from the success of these methods in general NLP (see for example Zhou et al (2005)). Statistical machine learning has also been applied to clinical text, but its use has generally been limited to entity recognition. The Mayo Clinic text analysis system (Pakhomov et al., 2005), for example, uses a combination of dictionary lookup and a Naïve Bayes classifier to identify entities for information retrieval applications. To the best of our knowledge, statistical methods have not been previously applied to extraction of clinical relationships from text.

This paper describes experiments in the statistical machine learning of relationships from a novel text type: oncology narratives. The set of relationships extracted are considered to be of interest for clinical and research applications down line of IE, such as querying to support clinical research. We apply Support Vector Machine (SVM) classifiers to learn these relationships. The classifiers are trained and evaluated using novel data: a gold standard corpus of clinical text, hand-annotated with semantic entities and relationships. In order to test the applicability of this method to the clinical domain, we train classifiers using a number of comparatively simple text features, and look at the contribution of these features to system performance. Clinically interesting relationships may span several sentences, and so we compare classifiers trained for both intra- and inter-sentential relationships (spanning one or more sentence boundaries). We also examine the influence of training corpus size on performance, as hand annotation of training data is the major expense in supervised machine learning.

2 Relationship Schema

Relationship	Argument 1	Argument 2
has_target	Investigation	Locus
	Intervention	Locus
has_finding	Investigation	Condition
	Investigation	Result
has_indication	Drug or device	Condition
	Intervention	Condition
	Investigation	Condition
has_location	Condition	Locus
negation_modifies	Negation modifier	Condition
laterality_modifies	Laterality modifier	Intervention
	Laterality modifier	Locus
sub-location_modifies	Sub-location modifier	Locus

Table 1: Relationship types and their argument type constraints.

The CLEF application extracts entities, relationships and modifiers from text. By *entity*, we mean some real-world thing, event or state referred to in the text: the drugs that are mentioned, the tests that were carried out, etc. *Modifiers* are words that qualify an entity in some way, referring e.g. to the laterality of an anatomical locus, or the negation of a condition ("no sign of inflammation"). Entities are connected to each other and to modifiers by *relationships*: e.g. linking a drug entity to the condition entity for which it is indicated, linking an investigation to its results, or linking a negating phrase to a condition.

The entities, modifiers, and relationships are described by both a formal XML schema, and by a set of detailed definitions. These were developed by a group of clinical experts through an iterative process, until acceptable agreement was reached. Entity types are mapped to types from the UMLS semantic network (Lindberg et al., 1993), each CLEF entity type covering several UMLS types. Relationship types are those that were felt necessary to capture the essential clinical dependencies between entities referred to in patient documents, and to support CLEF end user applications.

Each relationship type is constrained to exist between limited pairs of entity types. For example, the `has_location` relationship can only exist between a `Condition` entity and a `Locus` entity. Some relationships can exist between multiple type pairs. The full set of relationships and their argument type constraints are shown in Table 1. Examples of each relationship are given in Roberts et al (2008a).

Some of the relationships considered important by the clinical experts were not obvious without domain knowledge. For example,

> He is suffering from nausea and severe headaches. Dolasteron was prescribed.

Without domain knowledge, it is not clear that there is a `has_indication` relationship between the "Dolasteron" `Drug or device` entity and the "nausea" `Condition` entity. As in this example, many of this type of relationship are intra-sentential.

A single real-world entity may be referred to several times in the same text. Each of these co-referring expressions is a *mention* of the entity. The gold standard includes annotation of co-reference between different textual mentions of the same entity. For the work reported in this paper, however, co-reference is not considered. Each entity is assumed to have a single mention. Relationships between entities can be considered, by extension, as relationships between the single mentions of those entities. The implications of this are discussed further below.

3 Gold Standard Corpus

The schema and definitions were used to hand-annotate the entities and relationships in 77 oncology narratives, to provide a gold standard for system training and evaluation. Corpora of this size are typical in supervised machine learning, and reflect the expense of hand annotation. Narratives were carefully selected and annotated according to a best practice methodology, as described in Roberts et al (2008a). Narratives were annotated by two independent, clinically trained, annotators, and a consensus created by a third. We will refer to this corpus as C77.

Annotators were asked to first mark the mentions of entities and modifiers, and then to go through each of these in turn, deciding if any had relationships with mentions of other entities. Although the annotators were marking co-reference between mentions of the same entity, they were asked to ignore this with respect to relationship annotation. Both the annotation tool that they were using and their annotation guidelines, enforced the creation of relationships between mentions, and not between entities. The gold standard is thus analogous to the style of relationship extraction reported here, in which we extract relations between single mention entities, and do not consider co-reference. Annotators were further told that relationships could span multiple sentences, and that it was acceptable to use clinical domain knowledge to infer that a relationship existed between two mentions. Counts of all relationships annotated in C77 are shown in Table 2, sub-divided by the number of sentence boundaries spanned by a relationship.

4 Relationship Extraction

The system we have built uses the GATE NLP toolkit (Cunningham et al., 2002) [1]. The system is shown in Figure 1, and is described below.

Narratives are first pre-processed using standard GATE modules. Narratives were tokenised, sentences found with a regular expression-based sentence splitter, part-of-speech (POS) tagged, and morphological roots found for tokens. Each token was also labelled with a generalised POS tag, the first two characters of the full POS tag. This takes advantage of the Penn Treebank tagset used by GATE's POS tagger, in which related POS tags share the first two characters. For example, all six verb POS tags start with the letters "VB".

After pre-processing, mentions of entities within the text are annotated. In the experiments reported, we assume perfect entity recognition, as given by the entities in the human annotated gold standard

[1] We used a development build of GATE 4.0, downloadable from `http://gate.ac.uk`

	Sentence boundaries between arguments											**Total**
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**>9**	
has_finding	265	46	25	7	5	4	3	2	2	2	0	361
has_indication	139	85	35	32	14	11	6	4	5	5	12	348
has_location	360	4	1	1	1	1	1	0	0	0	4	373
has_target	122	14	4	2	2	4	3	1	0	1	0	153
laterality_modifies	128	0	0	0	0	0	0	0	0	0	0	128
negation_modifies	100	1	0	0	0	0	0	0	0	0	0	101
sub_location_modifies	76	0	0	0	0	0	0	0	0	0	0	76
Total	1190	150	65	42	22	20	13	7	7	8	16	1540
Cumulative total	1190	1340	1405	1447	1469	1489	1502	1509	1516	1524	1540	

Table 2: Count of relationships in 77 gold standard documents.

described above. Our results are therefore higher than would be expected in a system with automatic entity recognition. It is useful and usual to fix entity recognition in this way, to allow tuning specific to relationship extraction, and to allow the isolation of relation-specific problems. We accept, however, that ultimately, relation extraction does depend on the quality of entity recognition. The relation extraction described here is used as part of an operational IE system in which clinical entity recognition is performed by a combination of lexical lookup and supervised machine learning. We have described our entity extraction system elsewhere (Roberts et al., 2008b).

4.1 Classification

We treat clinical relationship extraction as a classification task, training classifiers to assign a relationship type to an *entity pair*. An entity pair is a pairing of entities that may or may not be the arguments of a relation. For a given document, we create all possible entity pairs within two constraints. First, entities that are paired must be within n sentences of each other. For all of the work reported here, unless stated, $n \leq 1$ (crossing 0 or 1 sentence boundaries). Second, we can constrain the entity pairs created by argument type (Rindflesch and Fiszman, 2003). For example, there is little point in creating an entity pair between a `Drug or device` entity and a `Result` entity, as no relationships, as specified by the schema, exist between entities of these types. Entity pairing is carried out by a GATE component developed specifically for clinical relationship extraction. In addition to pairing entities according to the above constraints, this component also assigns features to each pair that characterise its lexical and syntactic qualities (described further in Section 4.2).

Entity pairs correspond to classifier training and test instances. In classifier training, if an entity pair corresponds to the arguments of a relationship present in the gold standard, then it is assigned a class of that relationship type. If it does not correspond to such a relation, then it is assigned the class `null`. The classifier builds a model of these entity pair training instances, from their features. In classifier application, entity pairs are created from unseen text, under the above constraints. The classifier assigns one of our seven relationship types, or `null`, to each entity pair.

We use Support Vector machines (SVMs) as trainable classifiers, as these have proved to be robust and efficient for a range of NLP tasks, including relation extraction. We use an SVM implementation developed within our own group, and provided as part of the GATE toolkit. This is a variant on the original SVM algorithm, SVM with uneven margins, in which classification may be biased towards positive training examples. This is particularly suited to NLP applications, in which positive training examples are often rare. Full details of the classifier are given in Li et al (2005). We used the implementation "out of the box", with default parameters as determined in experiments with other data sets.

SVMs are binary classifiers: the multi-class problem of classifying entity pairs must therefore be mapped to a number of binary classification problems. There are several ways in which a multi-class problem can be recast as binary problems. The commonest are *one-against-one* in which one classifier is trained for every possible pair of classes, and *one-against-all* in which a classifier is trained for a binary decision between each class and all other classes, including null, combined. We have carried out extensive experiments (not reported here), with these two strategies, and have found little difference between them for our data. We have chosen to use one-against-all, as it needs fewer classifiers (for an n class problem, it needs n classifiers, as opposed to $\frac{(n-1)!}{2}$ for one-against-one).

The resultant class assignments by multiple binary classifiers must be post-processed to deal with ambiguity. In application to unseen text, it is possible that several classifiers assign different classes to an entity pair (test instance). To disambiguate these cases, the output of each one-against-all classifier is transformed into a probability, and the class with the highest probability is assigned. Re-casting the multi-class relation problem as a number of binary problems, and post-processing to resolve ambiguities, is handled by the GATE Learning API.

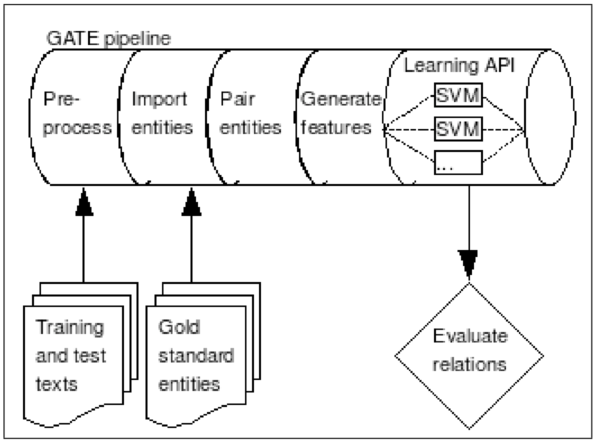

Figure 1: The relationship extraction system.

4.2 Features for Classification

The SVM classification model is built from lexical and syntactic features assigned to tokens and entity pairs prior to classification. We use features developed in part from those described in Zhou et al (2005) and Wang et al (2006). These features are split into 11 sets, as described in Table 3.

The tokN features are POS and surface string taken from a window of N tokens on each side of each paired entity's mention. For $N = 6$, this gives 48 features. The rationale behind these simple features is that there is useful information in the words surrounding two mentions, that helps determine any relationship between them. The gentokN features generalise tokN to use morphological root and generalised POS. The str features are a set of 14 surface string features, encoding the full surface strings of both entity mentions, their heads, their heads combined, the surface strings of the first, last and other tokens between the mentions, and of the two tokens immediately before and after the leftmost and rightmost mentions respectively. The pos, root, and genpos feature sets are similarly constructed from the POS tags, roots, and generalised POS tags of the entity mentions and their surrounding tokens. These four feature sets differ from tokN and gentokN, in that they provide more fine-grained information about the position of features relative to the paired entity mentions.

For the event feature set, the main entities were divided into events (Investigation and Intervention) and non-events (all others). Features record whether the entity pair consists of two events, two non-events, one of each, and whether there are any intervening events and non-events. This feature set gives similar information to atype (semantic types of arguments) and inter (intervening entities), but at a coarser level of typing.

5 Evaluation

We used a standard ten-fold cross validation methodology and standard evaluation metrics. Metrics are defined in terms of true positive, false positive and false negative matches between relationships in a system annotated *response* document and a gold standard *key* document. A response relationship is a true positive if a relationship of the same type, and with the exact same arguments, exists in the key. Corresponding definitions apply for false positive and false negative. Counts of these matches are used to calculate standard metrics of Recall (R), Precision (P) and $F1$ measure.

The metrics do not say how hard relationship extraction is. We therefore provide a comparison with Inter Annotator Agreement (IAA) scores from the gold standard. The IAA score gives the agreement between the two independent double annotators. It is equivalent to scoring one annotator against the other using the $F1$ metric. IAA scores are not directly comparable here, as relationship annotation is

Feature set	Size	Description
tokN	$8N$	Surface string and POS of tokens surrounding the arguments, windowed $-N$ to $+N$, $N = 6$ by default
gentokN	$8N$	Root and gerenalised POS of tokens surrounding the argument entities, windowed $-N$ to $+N$, $N = 6$ by default
atype	1	Concatenated semantic type of arguments, in arg1-arg2 order
dir	1	Direction: linear text order of the arguments (is arg1 before arg2, or vice versa?)
dist	2	Distance: absolute number of sentence and paragraph boundaries between arguments
str	14	Surface string features based on Zhou et al (2005), see text for full description
pos	14	POS features, as above
root	14	Root features, as above
genpos	14	Generalised POS features, as above
inter	11	Intervening mentions: numbers and types of intervening entity mentions between arguments
event	5	Events: are any of the arguments, or intevening entities, events?
allgen	96	All features in root and generalised POS forms, i.e. gentok6+atype+dir+dist+root+genpos+inter+event
notok	48	All except tokN features, others in string and POS forms, i.e. atype+dir+dist+str+pos+inter+event

Table 3: Feature sets used for learning relationships. The size of a set is the number of features in that set.

a slightly different task for the human annotators. The relationship extraction system is given entities, and finds relationships between them. Human annotators must find both the entities and the relationships. Therefore, were one human annotator to fail to find a particular entity, they could never find relationships with that entity. The raw IAA score does not take this into account: if an annotator fails to find an entity, then they will also be penalised for all relationships with that entity. We therefore give a Corrected IAA, CIAA, in which annotators are only compared on those relations for which they have both found the entities involved. Both forms of IAA are shown in Table 4. It is clear that it is hard for annotators to reach agreement on relationships, and that this is compounded massively by lack of perfect agreement on entities. Note that the gold standard used in training and evaluation reflects a further consensus annotation, to correct this poor agreement.

6 Results

6.1 Feature Selection

The first group of experiments reported looks at the performance of relation extraction with various feature sets. We followed an additive strategy for feature selection. Starting with basic features, we added further features one set at a time. We measured the performance of the resulting classifier each time we added a new feature set. Results are shown in Table 4. The initial classifier used a `tok6+atype` feature set. Addition of both `dir` and `dist` features give significant improvements in all metrics, of around 10% $F1$ overall, in each case. This suggests that the linear text order of arguments, and whether relations are intra- or inter-sentential is important to classification. Addition of the `str` features also give good improvement in most metrics, again 10% $F1$ overall. Addition of part-of-speech information, in the form of `pos` features, however, leads to a drop in some metrics, overall $F1$ dropping by 1%. Unexpectedly, POS seems to provide little extra information above that in surface string. Errors in POS tagging cannot be dismissed, and could be the cause of this. The existence of intervening entities, as coded in feature set `inter`, provides a small benefit. The inclusion of information about events, in the `event` feature set, is less clear-cut.

We were interested to see if generalising features could improve performance, as this had benefited our previous work in entity extraction. We replaced all surface string features with their root form, and POS features with their generalised POS form. This gave the results shown in column `allgen`. Results are not clear cut, in some cases better and in some worse than the previous best. Overall, there is no difference in $F1$. There is a slight increase in overall recall, and a corresponding drop in precision — as might be expected.

Both the `tokN`, and the `str` and `pos` feature sets provide surface string and POS information about tokens surrounding and between relationship arguments. The former gives features from a window around each argument. The latter two give a greater amount of positional information. Do these two provide enough information on their own, without the windowed features? To test this, we removed the `tokN` features from the full cumulative feature set, from column `+event`. Results are given in column

Relation	Metric	tok6+atype	+dir	+dist	+str	+pos	+inter	+event	allgen	notok	IAA	CIAA
has_finding	P	44	49	58	63	62	64	65	63	63		
	R	39	63	78	80	80	81	81	82	82		
	F1	39	54	66	70	69	71	72	71	71	46	80
has_indication	P	37	23	38	42	40	41	42	37	44		
	R	14	14	46	44	44	47	47	45	47		
	F1	18	16	39	39	38	41	42	38	41	26	50
has_location	P	36	36	50	68	71	72	72	73	73		
	R	28	28	74	79	79	81	81	83	83		
	F1	30	30	58	72	74	76	75	77	76	55	80
has_target	P	9	9	32	63	57	60	62	60	59		
	R	11	11	51	68	67	67	66	68	68		
	F1	9	9	38	64	60	63	63	63	62	42	63
laterality_modifies	P	21	38	73	84	83	84	84	86	86		
	R	9	55	82	89	86	88	88	87	89		
	F1	12	44	76	85	83	84	84	84	85	73	94
negation_modifies	P	19	54	85	81	80	79	79	77	81		
	R	12	82	97	98	93	92	93	93	93		
	F1	13	63	89	88	85	84	85	83	85	66	93
sub_location_modifies	P	2	2	55	88	86	86	88	88	87		
	R	1	1	62	94	92	95	95	95	95		
	F1	1	1	56	90	86	89	91	91	90	49	96
Overall	P	33	38	50	63	62	64	65	64	64		
	R	22	36	70	74	73	75	75	76	76		
	F1	26	37	58	68	67	69	69	69	70	47	75

Table 4: Variation in performance by feature set. Features sets are abbreviated as in Table 3. For the first seven columns, features were added cumulatively to each other. The next two columns, allgen and notok, are as described in Table 3. The final two columns give inter annotator agreement and corrected inter annotator agreement, for comparison.

notok. There is no clear change in performance, some relationships improving, and some worsening. Overall, there is a 1% improvement in $F1$.

It appears that the bulk of performance is attained through entity type and distance features, with some contribution from positional surface string information. Performance is between 1% and 9% lower than CIAA for the same relationship, with a best overall $F1$ of 70%, compared to a CIAA of 75%.

6.2 Sentences Spanned

Table 2 shows that although most relationships are intra-sentential, 23% are inter-sentential, 10% of all relationships being between arguments in adjacent sentences. If we consider a relationship to cross n sentence boundaries, then the classifiers described in the previous section were all trained on relationships crossing $n \leq 1$ sentence boundaries, i.e. with arguments in the same or adjacent sentences. What effect does including more distant relationships have on performance? We trained classifiers on only intra-sentential relationships, and on relationships spanning up to n sentence boundaries, for $n \in \{1...5\}$. We also trained a classifier on relationships with $1 \leq n \leq 5$, comprising 85% of all inter-sentential relationships. In each case, the cumulative feature set +event from Table 4 was used. Results are shown in Table 5. It is clear from the results that the feature sets used do not perform well on inter-sentential relationships. There is a 6% drop in overall $F1$ when including relationships with $n = 1$ together with $n < 1$. Performance continues to drop as more inter-sentential relationships are included, and is very poor for just inter-sentential relationships.

A preliminary error analysis suggests that the more distant relationship arguments are from each other, the more likely clinical knowledge is required to extract the relationship. This raises additional difficulties for extraction, which the simple features described here are unable to address.

6.3 Size of Training Corpus

The provision of sufficient training data for supervised learning algorithms is a limitation on their use. We examined the effect of training corpus size on relationship extraction. The C77 corpus, compris-

		Number of sentence boundaries between arguments									
		inter-	intra-	inter- and intra-sentential					Corpus size		
Relation	**Metric**	$1 \leq n \leq 5$	$n < 1$	$n \leq 1$	$n \leq 2$	$n \leq 3$	$n \leq 4$	$n \leq 5$	**C25**	**C50**	**C77**
has_finding	**P**	24	68	65	62	60	61	61	66	63	65
	R	18	89	81	79	78	78	77	74	74	81
	F1	18	76	72	69	67	68	67	67	67	72
has_indication	**P**	18	49	42	42	36	32	30	22	25	42
	R	17	59	47	42	42	39	38	30	31	47
	F1	16	51	42	39	37	34	33	23	25	42
has_location	**P**	0	74	72	73	72	72	72	72	71	72
	R	0	83	81	81	81	82	82	76	80	81
	F1	0	77	75	76	75	76	76	73	74	75
has_target	**P**	3	64	62	59	60	59	58	65	49	62
	R	1	75	66	64	62	61	61	60	65	66
	F1	2	68	63	61	60	60	59	59	54	63
laterality_modifies	**P**	0	86	84	86	86	86	87	77	78	84
	R	0	89	88	88	88	87	88	69	68	88
	F1	0	85	84	85	86	85	86	72	69	84
negation_modifies	**P**	0	80	79	79	80	80	80	78	79	79
	R	0	94	93	91	93	93	93	80	93	93
	F1	0	86	85	84	85	86	85	78	84	85
sub_location_modifies	**P**	0	89	88	88	89	89	89	64	91	88
	R	0	95	95	95	95	95	95	64	85	95
	F1	0	91	91	91	91	91	91	64	86	91
Overall	**P**	22	69	65	64	62	61	60	62	63	65
	R	17	83	75	73	71	70	70	65	71	75
	F1	19	75	69	68	66	65	65	63	66	69

Table 5: Variation in performance, by number of sentence boundaries (n), and by training corpus size.

ing 77 narratives and used in the previous experiments, was subsetted to give corpora of 25 and 50 narratives, which will be referred to as C25 and C50 respectively. We trained two further classifiers on these new corpora. Again, the cumulative feature set `+event` from Table 4 was used. Results are shown in Table 5. Overall, performance improves as training corpus size increases ($F1$ rising from 63% to 69%). We were struck however, by the fact that increasing from 50 to 77 documents has little effect on a few relationships (`negation_modifies` and `has_location`). It may well be that the amount of training data required has plateaued for those relationships.

7 Conclusion

We have shown that it is possible to extract clinical relationships from text, using shallow features, and supervised statistical machine learning. Judging from poor inter annotator agreement, the task is hard. Our system achieves a reasonable performance, with an overall $F1$ just 5% below a corrected inter annotator agreement. This performance is reached largely by using features of the text that encode entity type, distance between arguments, and some surface string information. Performance does, however, vary with the number of sentences spanned by the relationships. Learning inter-sentential relationships does not seem amenable to this approach, and may require the use of domain knowledge.

A major concern when using supervised learning algorithms is the expense and availability of training data. We have shown that while this concern is justified in some cases, larger training corpora may not improve performance for all relationships.

The technology used has proved scalable. The full CLEF IE system, including automatic entity recognition, is able to process a document in subsecond time on a commodity workstation. We have used the system to extract 6 million relations from over half a million patient documents, for use in downstream CLEF applications (Roberts et al., 2008a). Our future work on relationship extraction in CLEF includes integration of a dependency parse into the feature set, further analysis to determine what knowledge may be required to learn intersentential relations, and integration of relationship extraction with a co-reference algorithm.

Availability All of the software described here is open source and can be downloaded as part of GATE, with the exception of the entity pairing component, which will be released shortly. We are currently preparing a UK research ethics committee application, requesting permission to release our annotated corpus.

Acknowledgements

CLEF is funded by the UK Medical Research Council. We would like to thank the Royal Marsden Hospital for providing the corpus, and our clinical partners in CLEF for assistance in developing the schema, and for gold standard annotation.

References

H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. 2002. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*, pages 168–175, Philadelphia, PA, USA, July.

C. Friedman, P. Alderson, J. Austin, J. Cimino, and S. Johnson. 1994. A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association*, 1(2):161–174, March.

C. Grover, B. Haddow, E. Klein, M. Matthews, L. Nielsen, R. Tobin, and X. Wang. 2007. Adapting a relation extraction pipeline for the BioCreAtIvE II task. In *Proceedings of the BioCreAtIvE II Workshop 2007*, Madrid, Spain.

U. Hahn, M. Romacker, and S. Schulz. 2002. MEDSYNDIKATE — a natural language system for the extraction of medical information from findings reports. *International Journal of Medical Informatics*, 67(1–3):63–74, December.

Y. Li, K. Bontcheva, and H. Cunningham. 2005. SVM based learning system for information extraction. In *Deterministic and statistical methods in machine learning: first international workshop*, number 3635 in Lecture Notes in Computer Science, pages 319–339. Springer.

D. Lindberg, B. Humphreys, and A. McCray. 1993. The Unified Medical Language System. *Methods of Information in Medicine*, 32(4):281–291.

Y. Lussier, T. Borlawsky, D. Rappaport, Y. Liu, and C. Friedman. 2006. PhenoGO: Assigning phenotypic context to Gene Ontology annotations with natural language processing. In *Biocomputing 2006, Proceedings of the Pacific Symposium*, pages 64–75, Hawaii, USA, January.

S. Pakhomov, J. Buntrock, and P. Duffy. 2005. High throughput modularized NLP system for clinical text. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), interactive poster and demonstration sessions*, pages 25–28, Ann Arbor, MI, USA, June.

A. Rector, J. Rogers, A. Taweel, D. Ingram, D. Kalra, J. Milan, P. Singleton, R. Gaizauskas, M. Hepple, D. Scott, and R. Power. 2003. CLEF — joining up healthcare with clinical and post-genomic research. In *Proceedings of UK e-Science All Hands Meeting 2003*, pages 264–267, Nottingham, UK.

T. Rindflesch and M. Fiszman. 2003. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *Journal of Biomedical Informatics*, 36(6):462–477.

A. Roberts, R. Gaizauskas, M. Hepple, G. Demetriou, Y. Guo, A. Setzer, and I. Roberts. 2008a. Semantic annotation of clinical text: The CLEF corpus. In *Proceedings of Building and evaluating resources for biomedical text mining: workshop at LREC 2008*, Marrakech, Morocco, May. In press.

A. Roberts, R. Gaizauskas, M. Hepple, and Y. Guo. 2008b. Combining terminology resources and statistical methods for entity recognition: an evaluation. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC 2008*, Marrakech, Morocco, May. In press.

N. Sager, M. Lyman, C. Bucknall, N. Nhan, and L. Tick. 1994. Natural language processing and the representation of clinical data. *Journal of the American Medical Informatics Association*, 1(2):142–160, March-April.

T. Wang, Y. Li, K. Bontcheva, H. Cunningham, and J. Wang. 2006. Automatic extraction of hierarchical relations from text. In *The Semantic Web: Research and Applications. 3rd European Semantic Web Conference, ESWC 2006*, number 4011 in Lecture Notes in Computer Science, pages 215–229. Springer.

G. Zhou, J. Su, J. Zhang, and M. Zhang. 2005. Exploring Various Knowledge in Relation Extraction. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 427–434, Ann Arbor, MI, USA, June.

P. Zweigenbaum, B. Bachimont, J. Bouaud, J. Charlet, and J-F. Boisvieux. 1995. A multi-lingual architecture for building a normalised conceptual representation from medical language. In *Proceedings of the Annual Symposium on Computer Applications in Medical Care*, pages 357–361, New York, NY, USA.