Using Automated Feature Optimisation to Create an Adaptable Relation Extraction System

Barry Haddow
School of Informatics, University of Edinburgh,
2 Buccleuch Place, Edinburgh, Scotland, EH8 9LW
bhaddow@inf.ed.ac.uk

Abstract

An adaptable relation extraction system for the biomedical domain is presented. The system makes use of a large set of contextual and shallow syntactic features, which can be automatically optimised for each relation type. The system is tested on three different relation types; protein-protein interactions, tissue expression relations and fragment to parent protein relations.

1 Introduction

In biomedical information extraction, research in named entity recognition (NER) and relation extraction (RE) has tended to focus on the extracting proteins and their interactions, with less thought given to how to adapt such systems to other entities and relations of biomedical interest. This is especially true for RE, where there is very little work on relations other than protein-protein interactions. Nevertheless, in order to create applications of use to biologists such as curation assistants and improved information extraction and retrieval systems it will be necessary to treat a broader range of semantic relations. The recent release of the Genia event corpus (Kim et al., 2008) will help to drive this research.

The aim of this paper is to address the problem of how to create an RE system, which can be adapted to different biomedical RE problems with a minimum of manual intervention. Since this paper focuses on relation extraction, it will be assumed that the named entities are given, in other words the human annotated entities are used in all experiments. The approach taken to RE is to treat it as a supervised classification problem on relation candidates, using a large collection of shallow syntactic and contextual features. Relation candidates are pairs of entities, picked out using an appropriate candidate generation strategy. The use of shallow (as opposed to deep) syntactic features means that the system can rely on relatively robust linguistic tools such as part-of-speech taggers and chunkers, rather than more brittle and less widely available tools such as parsers. The difficulty with feature-based methods is, however, how to select the best performing feature set, as simply adding all possible features does not necessarily give the best results (Guyon and Elisseeff, 2003). The approach taken here is to implement a large feature set and then use a greedy search to explore the feature set and select the best subset of features. This method of feature set optimisation is not new (for example, it was applied by one team (Ganchev et al., 2007) on the BioCreative II Gene Mention task), but in this work a comparison of search starting points and feature groupings will be presented.

All RE systems require a human-annotated corpus for testing, and since a supervised machine learning approach is employed, a corpus is also required for training the system. The experiments described in this paper make use of the ITI TXM corpora (Alex et al., 2008), which include the PPI corpus addressing protein-protein interactions, and the TE corpus addressing tissue expression. Both corpora consist of approximately 200 full-text biomedical research papers annotated with entities, normalisations of entities to standard databases, relations, and with enriched information added to the relations. Only the entities and relations will be considered here.

This paper is organised as follows: after reviewing related work in the following section, the RE system is described in Section 3, including a description of the corpora, the relation candidate extraction strategies, the features employed, the feature optimisation methods and the evaluation method. In Section 4 the results of the optimisation experiments are presented and discussed, with some concluding remarks in Section 5.

BioNLP 2008: Current Trends in Biomedical Natural Language Processing, pages 19–27,
Columbus, Ohio, USA, June 2008. ©2008 Association for Computational Linguistics

2 Related Work

Recent interest in the extraction of protein-protein interactions has been given added impetus by shared tasks such as the Language Learning in Logic (Cussens and Nédellec, 2005), and the BioCreative II Interaction Pairs Subtask (Krallinger et al., 2008). It should be noted that the latter task, rather than being concerned with the extraction of specific interaction relation mentions, required systems to list the (curatable) interactions at a document level. Many teams, however, extracted the interaction mentions as a first step and then processed these to give the document level list of curatable interactions.

The extraction of protein-protein interactions has also been helped by the availability of annotated corpora, such as AIMed (Bunescu et al., 2005), which consists of around 1000 Medline abstracts annotated with proteins and their interactions. In common with the LLL corpus, the AIMed corpus only contains intra-sentential relations, and is somewhat smaller than the corpus used in the current work. In addition to the work by the corpus creators (Bunescu and Mooney, 2007), other authors have achieved good results on AIMed by making use of dependency parses in different ways (Erkan et al., 2007; Katrenko and Adriaans, 2006). It is not clear, however, how well these techniques would transfer to other, similar, RE problems, and how much work would be involved in tuning the systems for a new problem.

Supervised learning based on shallow syntactic features has also been applied to the biomedical domain, again focusing on protein-protein interactions (Nielsen, 2006; Giuliano et al., 2006). A systematic exploration of a set of such features for protein-protein interaction extraction was recently provided by Jiang and Zhai (2007), who also used features derived from the Collins parser. They did not, however, experiment with the automated optimisation of the feature sets. In the news domain, the best reported results on the ACE dataset[1] have been achieved by a composite kernel which depends partially on a full parse, and partially on a collection of shallow syntactic features (Zhou et al., 2007).

Aside from protein-protein interactions, there has been little work directed at other types of relations in the biomedical domain. Recent corpus annotation projects such as Genia (Kim et al., 2008) and BioInfer (Pyysalo et al., 2007) include multiple types of relations, however many of the relation types are represented in fairly small quantities. In earlier work (Skounakis et al., 2003), the extraction of cell localisation relations was studied using an automatically created corpus.

3 Methods

3.1 Corpora

The ITI TXM corpora contain annotations related to protein-protein interactions (in the PPI corpus), and annotations related to tissue expression experiments (in the TE corpus). Each corpus consists of biomedical research articles, selected from PubMed and PubMedCentral either because they contain experimentally proven protein-protein interactions (for the PPI corpus), or because they contain tissue expression experiments (for the TE corpus).

The articles were annotated by a team of qualified biologists. The annotations consisted of entities (Table 1), normalisations of selected entities to standard databases, relations (Table 2) and enrichment of relations with additional information of interest to curators. For each corpus, the entities marked were those involved in the relation which formed the principal focus of that corpus (either PPI or TE), and those which could affect this relation. In the TE corpus, TE relations were marked when the text stated that a particular gene or gene product was present or absent in a particular tissue, whilst PPI relations were marked whenever a statement (positive or negative) was made about the interaction of a pair of Proteins, Mutants, Fragments, Complexes or Fusions. In addition, both corpora were annotated with FRAG relations which connect Fragments and Mutants with their parent Proteins.

Corpus	Entities
PPI	CellLine, Complex, DrugCompound, ExperimentalMethod, Fragment, Fusion, Modification, Mutant, Protein
TE	Complex, DevelopmentalStage, Disease, DrugCompound, ExperimentalMethod, Fragment, Fusion, GOMOP, Gene, Mutant, Protein, Tissue, mRNAcDNA

Table 1: The entity types in the TE and PPI corpora. Note that *GOMOP* stands for *"Gene or mRNAcDNA or Protein"* and was used when the annotators felt the author was using the term in an ambiguous way.

In order to monitor annotation quality, and to measure of the difficulty of the task, some documents were multiply annotated. The counts of the numbers of unique documents in each section, together with the numbers of annotated documents are shown in Table 3. Note that the multiply annotated documents were not reconciled, but the multiple copies were included in the corpus. Each corpus was split into three sections – TRAIN, DEVTEST and TEST – with the first two sections being used for system development, and the last reserved for final testing.

[1] http://www.nist.gov/speech/tests/ace/

Corpus	Relation type	Entity 1 Types	Entity 2 Types	Count
PPI	PPI	Protein, Fusion, Mutant, Fragment or Complex	Protein, Fusion, Mutant, Fragment or Complex	11,523
	FRAG	Protein	Mutant or Fragment	16,002
TE	TE	Gene, Protein, mRNAcDNA, GOMOP, Fusion, Mutant, Complex or Fragment	Tissue	12,426
	FRAG	Protein	Mutant or Fragment	4,735

Table 2: Relation types in each corpus.

Corpus	Segment	Unique Documents	Annotated Documents
PPI	TRAIN	133	221
	DEVTEST	39	58
	TRAIN	45	57
TE	TRAIN	151	221
	DEVTEST	41	48
	TEST	46	59

Table 3: Counts of documents and annotations in each corpus.

Corpus	Relation	Intra	Inter
PPI	PPI	10,607(92.1%)	916(7.9%)
	FRAG	10,176(63.6%)	5,826(36.4%)
TE	TE	10,356(83.3%)	2,070(16.7%)
	FRAG	3,335(70.4%)	1,400(29.6%)

Table 4: Counts of inter and intra-sentential relations.

Annotators were permitted to mark relations between entities in the same sentence (*intra-sentential*), or between entities in different sentences (*inter-sentential*). The majority of relations were intra-sentential, with FRAG relations showing the highest proportion of inter-sententials. Table 4 shows the counts of inter/intra-sentential relations of each type.

Some examples of each type of relation will now be presented. The first example is from PubMed 16436664, and is a TE relation:

> Our recent observations that $\langle \alpha\text{v}\beta 5\rangle_1$ is upregulated in $\langle$scleroderma fibroblasts$\rangle_1$ and that the transient overexpression of $\alpha\text{v}\beta 5$ increases the human $\langle \alpha 2(\text{I})$ collagen$\rangle_2$ gene expression in normal $\langle$fibroblasts$\rangle_2$ ...

There are two different TE relations in this sentence fragment, indicated by the numerical subscripts; the first connects a Tissue and a Complex, and the second connects a Tissue with a Gene. Another example from the same paper shows a FRAG relation.

> Because $\langle \beta 5\rangle_1$ has a $\langle$cytoplasmic domain$\rangle_1$ highly homologous to that of $\beta 6$-subunit, 42 we made a hypothesis that $\alpha\text{v}\beta 5$ activates SLC by the nonproteolytic pathway.

The annotators could also mark negative TE and PPI relations, as shown in the following example of a PPI relation taken from PubMedCentral 1075921.

> It was also previously reported that two truncated versions of $\langle$p53$\rangle_{1,2}$, consisting of residues $\langle$2-45$\rangle_{1,3}$ and $\langle$46-71$\rangle_{2,4}$, do not bind $\langle$hRPA70$\rangle_{3,4}$ (47)

Here the PPI relations connect the two Fragments ("2-45" and "46-71") to the Protein "hRPA70", whilst FRAG relations connect the Fragments with their parent Protein "p53".

In contrast with the straightforward intra-sentential relations shown above, the following (from PubMed 16399077) is an example of an inter-sentential TE relation (only the related entities are shown).

> To test whether SPE can activate Toll signaling, we expressed activated SPE in $\langle$S2 cells$\rangle_1$ and in flies, and we then assayed the expression of the gene for Drosomycin (Drs), an antifungal peptide known to be induced by Toll signaling in response to microbial infection (Lemaitre et al., 1996). In both cases, $\langle$Drs$\rangle_2$ expression was significantly induced in the absence of infection,

In this example, the annotator has connected a Tissue on the first sentence, with an mRNAcDNA in the second.

The multiply annotated documents in the corpus were used to calculate the inter-annotator agreement (IAA), by scoring different versions of the annotation of the same document against each other. For each corresponding pair of annotations, one annotator was selected as the "gold", and the other annotator scored against the first using precision, recall and F_1 on relations. Only relations where both annotators agreed on the participating entities were considered. The scores for each annotated document pair were then micro-averaged (where each example

Corpus	Type	Intra	Inter	All
PPI	PPI	69.7	41.1	67.0
	FRAG	90.5	73.9	84.6
	All	78.7	67.3	76.1
TE	TE	72.8	59.4	70.1
	FRAG	89.7	69.0	84.0
	All	77.4	62.7	74.1

Table 5: IAA for relation annotation, split by inter- and intra-sentential

is given equal weight) to produce overall IAA scores for the corpus, shown in Table 5.

The main observations from Table 5 are that TE and PPI relations are harder to annotate than FRAG relations, and that inter-sentential are harder than intra-sentential. In particular, the IAA for intra-sentential FRAG relations is very high, probably because many of these are very straightforward constructions such as "Fragment of Protein". Inter-sentential relations are often less clear as they involve linking information between several sentences, for example using coreferences.

Both corpora were pre-processed before RE was applied. The pre-processing involved tokenisation, sentence boundary detection, lemmatising. part-of-speech tagging, head word detection and chunking. The part-of-speech tagging uses the Curran & Clark maximum entropy Markov model tagger (Curran and Clark, 2003) trained on MedPost data (Smith et al., 2004), whilst the other preprocessing stages are all rule-based. The tokenisation, sentence boundary detection, head word identification and chunking components were implemented with the LT-XML2 tools (Grover and Tobin, 2006), and the lemmatisation used `morpha` (Minnen et al., 2000).

3.2 The Relation Extraction System

Relation extraction is treated a classification problem, by generating candidate relations, and classifying them as either *true* or *false*. In the optimisation experiments described in this paper, Zhang Le's maximum entropy (MAXENT) classifier[2] was used, since its performance was very competitive and its fast training time permitted extensive feature experimentation. The Gaussian prior was set to 0.1, and the maximum training iterations to 100. In order to assess the performance of the final system, MAXENT was compared with support vector machines (SVM) using the SVM^{light} toolkit (Joachims, 1999). Since both the classifiers assign a confidence to each prediction, a varying threshold can be applied to the output of the classifier to provide a precision-recall tradeoff.

Candidate relations were generated by considering entity pairs of the appropriate type, taking into account the distance between the entities. It was thought that inter-sentential and intra-sentential relations would require different feature sets and different models, so inter- and intra-sentential candidates were generated separately. For intra-sentential relations, all entity pairs of the appropriate type (as in Table 2) in the same sentence were permitted as candidates, with the sole exclusion being that any entities contained in a Fusion entity were not allowed to participate in candidate TE relations. This restriction was in place in the annotation guidelines, so no such relations were annotated. For intra-sentential relations in the training data, around 25-30% of the candidate relations are actual relations.

Generating inter-sentential candidates is more problematic, as measures must be taken to limit the number of candidates. Inter-sentential FRAG candidates are restricted to a distance of no more than 5 sentences, whilst inter-sentential PPI and TE candidates are restricted to participants in adjacent sentences. Inter-sentential RE is performed after intra-sentential RE, so the candidate generation strategy has access to the annotated intra-sentential relations (in training) and the predicted intra-sentential relations (in testing). For TE and PPI, candidates are only created for those entities not already in a relation, and for FRAG candidates are only created if the Mutant or Fragment is not already in a relation. Furthermore, for FRAG relations, if there is more than one Protein instance with the same lexical form in the 5 sentence window, then a candidate relation is only created between a given Fragment/Mutant and the nearest occurrence of this Protein. For inter-sentential FRAG relations, around 20% of the candidates are actual relations, however for TE and PPI, only about 1% of the candidate relations are actual relations.

3.3 Features

Each candidate relation is mapped to a feature representation, where the features are binary or real-valued functions of relations. The majority of the features are binary, although these are actually special cases of real-valued functions, taking values 0 or 1. A feature representation of a relation is normally written as a sequence of strings, each corresponding to a different feature, and the presence or absence of a binary feature indicating whether it is on or off. In order that the relation extractor could be applied to different problems and optimised, a large number of features were implemented, with the intention that the feature space could be automatically searched to find the best subset.

[2] `http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html`

Features are normally grouped into *feature templates* and, as is common in the literature, the feature templates may also be referred to as *features.* For instance, a feature template may be “the token to the right of the second entity in the relation”, which then gives rise to a set of boolean features with the prefix `ctxt-w-rf1-`. One such feature in this set is the feature which indicates that the token to the right of the second entity is “the”, i.e. `ctxt-w-rf1-the`. The feature templates are then collected into *feature groups*, such as “context features”, which are really just a convenient way of conceptualising, implementing and managing the features, and do not necessarily reflect any common behaviour amongst the features in a group.

The following is a comprehensive list of the feature implemented in the RE system with features listed by group, and the possible options for each feature group given. The options are used to turn on or off feature templates in the group, or change templates, and may be boolean or numerical. The nature of the options will be important in the feature exploration experiments since they influence the type of search operations which may be used to explore the feature space. The features are virtually all domain independent, except for perhaps the SignSlashSign feature which is specific to TE. The RelationKeywordFeature can easily be ported to a new domain by generating a list of keywords appropriate for the given relation.

In the feature group descriptions which follow, the term “participants” refers to the entities within the candidate relation. Some of the features make use of the “vlw backoff”, which for a given token is defined as the verb stem, backing off to the lemma if that is not available, backing off to the token itself.

Chunk This group has three optional templates; one which adds the concatenated sequence of chunk types between the participants, and two templates which add the count of chunks between the participants as binary and numeric features, respectively. So if the chunk count is, for example 4, the binary feature would be `chunk-bwcount-4` and the numeric feature would have name `chunk-bwcount` and value 4.

EntitiesBetween This has templates to indicate the type and relative position of the entities between the two participants. For TE relations, only Tissue entities are considered, whilst for other relations only Proteins are considered.

Entity Features derived from the participating entities are added by the templates in this group, which has options to turn on the entity’s text, class and bigrams of these. There is also a feature template which adds all words in the entities as separate features, and one that adds all words in the second entity only, plus options to add features which indicate when the two entities have the same textual form, or when one is a substring of the other.

EntityContext The entity context can include tokens, part-of-speech tags, chunk tags and vlw backoffs, each within window sizes determined by numerical options. A further option can switch on a template which adds the concatenation of all vlw backoffs in the context, on either side of each entity, and there is also an option to convert all tokens and vlw backoffs to lower case before creating the features.

EntityDistance Options on this group allow the addition of the token distance and sentence distance between the entities, as numeric or binary features. There is also an option to add a coarse three-way classification of the token distance.

EntityFrequency Counts are made of the number of occurrences of each entity surface form in the document, limited to Tissue entities for TE relations, and Proteins for FRAG and PPI relations. The only option for this feature group adds a template which generates a binary feature indicating the frequency rank of the participants’ surface forms in the document.

EntityPattern The entity pattern for a given intrasentential candidate relation shows how its participants lie with respect to the other entities in the sentence. The pattern is a concatenated sequence of the entity types in the sentence, with the participants in upper case and other entities in lower case. Only entity types which are valid participants in the relation in question are included. For example `protein-PROTEIN-TISSUE` would indicate a relation between a Protein and a Tissue, with another Protein occurring first in the sentence. Options in this group add the patterm, the total number of entities in the sentence, and the numbers of entities for each type.

Frame The frame is the concatenation of the tokens between the two participants. Two boolean options on this group specify whether or not to include the token concatenation, and whether or not to include the part-of-speech concatenation. A further numeric option is used to limit the maximum frame length; when this is set to a non-zero value longer frames are discarded.

HeadWord All the headwords of the chunks in the sentences containing and between the participants are listed and used to construct the features in this group. Options specify whether to include head nouns and/or head verbs, and whether to convert the headwords to lower case or replace them by their vlw backoffs. A further option allows an additional marker to be added to each headword feature to indicate whether it is before, between or after the participants.

NestedEntity This feature indicates whether the participants are contained in other entities, or in each other. The first option adds a feature template which indicates which type of entity containing the two participants, if they are both contained. The second option adds a feature to indicate whether one of the entities is contained within the other, and the third adds a feature to indicate whether or not there is any whitespace between the two entities.

Ngram Three options specify what type of ngrams to add; whether to add unigrams of the tokens in the sentences containing the participants, whether to add bigrams of the same tokens, and whether to add cross-bigrams, which are bigrams of tokens before and between the participants, and of tokens between and after the participants. Additional options specify whether to convert tokens to vlw backoff or lower case and whether to replace all sequences of digits by "0". Further options can be used to indicate that only ngrams in between the participants should be added, that each ngram feature should be marked as before, between or after, or that all entities should replaced in the text by their type.

RelationKeyword Relation keywords are terms annotated as relation indicators for PPI and TE, and linked to relations. For PPI they are interaction words, and for TE they are expression level words. Keywords are matched from a list generated during training and there are feature options to match these keywords before, between and after the participants, and to add templates for the existence of a keyword, the text of the keyword, and whether or not it is a head word.

RelativeEntityPosition The only option on this group specifies whether or not to sort the participant entities, alphabetically by entity type. Binary features are added indicating whether the first entity in the candidate relation is the first in the document, whether it is the second, whether the participants overlap or whether they coincide.

SignSlashSign This group is only used for TE relations and is designed to detected the presence of indicators like +/+ and −/+ in the sentence(s) containing the relation. Options allow the existence and type of the one of these expressions to be indicated, and also its position relative to the participants, and whether it is adjacent to one of the participants.

3.4 Optimisation

Feature selection methods include *wrapper* methods where feature sets are assessed according to their effectiveness for a given learner, and *filter* methods where features are removed using some criterion before being passed to the learner (Guyon and Elisseeff, 2003). In building the RE system, it was found that filter methods did not work well, probably due to the large number of interactions between the features, so a wrapper optimisation method was employed, consisting of greedy search through the space of possible feature sets.

In the greedy search method, an initial feature set is selected and a model trained on the TRAIN set and tested on the DEVTEST set. A series of search operators (see below) are applied to the feature set to produce a list of proposed new feature sets, one corresponding to each operator, and the new feature sets are tested in the same way. If any of these new feature sets produces better results than the original initial set, then the best set replaces the initial feature set and the process is iterated. The greedy search terminates when none of the search operators leads to an improvement. Three types of search operators are used in the greedy search, defined in terms of the feature set structure described in Section 3.3:

1. The deletion of a feature group.
2. The increase or decrease of a numerical option on a feature group (e.g. context size), where the size of the change is not greater than 2.
3. The flipping of a boolean option on a feature group.

In theory search operators which add or remove individual features could be used, but due to the large number of features the use of such operators is not practical. In addition, it may have been possible to achieve more robust results using cross-validation rather than heldout testing, but that would also result in a large increase in search time.

3.5 Evaluation

In all RE experiments, the annotated entities were assumed as given so that only RE performance was being assessed. The performance was measured using *precision-recall break-even point* (BEP), which is found by adjusting the decision boundary (threshold) of the classifier until the precision and recall are equal then taking the value of the F_1 at this threshold. The BEP has the advantage over F_1 that its definition is independent of the choice of threshold, but it can still be compared easily to the IAA and is based on the familiar concepts of precision and recall.

4 Results

Performance of the RE system on each of the four relation types was optimised using the greedy feature exploration method described in Section 3.4. Inter and intra-sentential relations were treated separately, with intra-sentential relation performance optimised first. The inter-sentential performance was then assessed using a "pipeline" consisting of the best intra-sentential relation extractor, and the inter-sentential system being optimised.

The greedy search experiments for intra-sentential relations used two different starting feature sets, an **all** set in which all features groups and options were switched on, and the context sizes in *EntityContext* were set to 3, and a **base** set which used just *Ngram* and *RelativeEntityPosition* features. The models were trained on TRAIN and scored on DEVTEST using BEP. In the calculation of BEP, all relations of the appropriate type were considered, including inter-sententials. The results of the greedy search on intra-sentential relations are shown in Table 6.

Corpus	Relation Type	Initial Features	Initial BEP	Final BEP
PPI	PPI	**base**	36.8	52.2
		all	51.6	53.4
	FRAG	**base**	49.2	56.0
		all	55.9	57.4
TE	TE	**base**	45.9	51.9
		all	50.6	53.8
	FRAG	**base**	53.7	62.7
		all	60.1	61.2

Table 6: Greedy search feature exploration for intra-sentential relations. Performance is measured on all relations, testing on DEVTEST.

For all relation types, the greedy search improves the performance over the **base** and **all** feature sets, usually reaching the highest performance when starting from **all**. Comparing the results in Table 6 with the IAA figures provided in Table 5 shows that the system performance is around 75-80% of IAA, with the lowest relative performances observed for FRAG relations. These relations include a higher proportion of inter-sententials, so systems which ignore inter-sententials suffer a larger loss in performance.

After choosing the best system for intra-sentential relations, the same greedy optimisation was performed on the inter-sentential relations using virtually the same initial feature sets. The only difference in the feature sets is that additional options are added to the *EntityDistance* feature to indicate the sentential distance between the entities. The result of the greedy search on the inter-sentential relations is shown in Table 7.

The inter-sentential relation optimisation is only really successful for the FRAG relations in the PPI corpus. For TE and PPI inter-sentential relations, the number of negative examples dwarfs the few positive examples making it very difficult for the machine learner. For FRAG relations in both corpora, some progress is made on the performance on inter-sentential relations (detailed breakdown not shown) but in the TE corpus this does not translate to an overall improvement in BEP. This is because the inter- and intra-sentential probabilities have quite different ranges for FRAG relations meaning that the threshold probabilities would have to be chosen separately to give the best F_1 score.

Corpus	Relation Type	Initial Features	Initial BEP	Final BEP
PPI	PPI	**base**	53.4	53.4
		all	53.4	53.4
	FRAG	**base**	59.6	62.2
		all	61.7	62.5
TE	TE	**base**	53.9	54.0
		all	53.9	54.0
	FRAG	**base**	60.4	62.8
		all	62.6	62.7

Table 7: Greedy search feature exploration for inter-sentential relations. Performance is measured on all relations, testing on DEVTEST.

The greedy search results just presented were based on a partitioning of the feature sets into groups which correspond to the way in which the features were implemented. Since the search operators apply at group granularity, and are not able to select features from within a group, the way in which the features are grouped is likely to have a bearing on the performance of the best system found by the algorithm. The next set of experiments investigates the effective the feature grouping by conducting greedy search with groups chosen randomly.

Corpus	Relation Type	Initial BEP	Final BEP	Ensemble BEP
PPI	PPI	51.1	52.9, 52.4, 52.7, 52.8, 52.6	52.5
	FRAG	55.7	56.3, 56.1, 56.1, 56.3, 56.4	56.3
TE	TE	51.4	52.0 , 51.8, 52.5, 51.9, 52.9	52.1
	FRAG	60.1	60.8, 60.5, 60.4, 60.7, 60.5	60.4

Table 8: Greedy search feature exploration with random feature groupings for intra-sentential relations. The initial feature set is a slightly modified **all** in each case, and the search was run 5 times, testing on DEVTEST. The ensemble system combines the 5 optimised feature sets using the geometric mean probability.

Using a variant of the **all** feature set where the context sizes in *EntityContext* were set to 5, a greedy search for the best performing system was implemented by first dividing the feature set randomly into 50 groups, and at each iteration testing the performance with each group added and removed in turn. The search was iterated until no further improvement in performance was obtained, where performance was measured using BEP. As for the previous greedy feature optimisations, the relation extractor was trained on TRAIN and tested on DEVTEST. The results for intra-sentential relations are shown in Table 8, where the experiment was repeated several times with different (randomly chosen) partitions. After performing the five random knockout searches of the feature space, an ensemble system was created for each relation type by training a system with each feature set and combining the five by taking the geometric mean of the probabilities. The performance of the ensemble system is shown in the final column of Table 8.

Comparing the results in Table 8 with the corresponding results for intra-sentential relations in Table 6, it can be seen that splitting the features into related groups works better than random groups. The ensemble does not improve on the individual scores, probably because the systems in the ensemble are not diverse enough (Dietterich, 2000)

To see how well the best feature sets generalise to unseen data, RE systems were trained on TRAIN and DEVTEST combined, and tested on TEST using different feature sets; the baseline sets (**base** and **all**), and the fully optimised set (**best**). In addition, to ensure that the greedy feature optimisation was not biasing the feature set towards the particular learner employed (i.e. MAXENT), systems were also trained and tested using SVM. The MAXENT system had its Gaussian prior optimised on the DEVTEST set, whilst SVM was found to work best with a linear kernel, and its cost factor was optimised on DEVTEST. The value of the decision function was used for thresholding the SVM model in order to calculate the BEP. The comparison of all systems on TEST is shown in Table 9.

Corpus	Relation Type	Learner	Feature Set		
			base	**all**	**best**
PPI	PPI	MAXENT	39.7	48.3	49.1
		SVM	39.6	49.2	**49.9**
PPI	FRAG	MAXENT	56.9	68.0	69.4
		SVM	54.9	68.2	**69.5**
TE	TE	MAXENT	39.0	47.9	46.8
		SVM	39.6	49.8	**50.1**
TE	FRAG	MAXENT	60.1	63.4	68.9
		SVM	59.7	67.7	**70.4**

Table 9: The performance of the system trained on TRAIN and DEVTEST, and tested on TEST. Performance is compared across the baseline feature sets (**base** and **all**) and the optimised feature set (**best**) using each classifier.

The results in Table 9 show that, in general, both classifiers perform better with the **all** feature set than with the **base** feature set, and best of all with the **best** feature set. The SVM classifier preserves this ordering throughout, and actually performs better than the MAXENT classifier overall, even though the features were optimised for MAXENT. For MAXENT, the **best** model outperforms **all** in three out of four cases, with the exception being TE.

5 Conclusions

It has been shown that a relation extraction system based on a supervised classifier and a large collection of shallow linguistic features can be applied to three different types of relations in two different biomedical corpora. Automated feature optimisation produced small gains in performance which were still apparent on a blind test set. Even though a wrapper method was used using a specific classifier (MAXENT), the feature set optimisations were still valid for an SVM classifier.

Since the greedy search through feature space is essentially a beam search with a beam size of one, it could be extended by using a larger beam-size, running the feature set comparisons in parallel to reduce total running time to a manageable size. Ad-hoc experiments have suggested that better results could be obtained by restarting the feature optimisation in different positions, indicating that local optima could be a problem, but a thorough investigation of the search space nature has been left for future work. Furthermore, the hyperparameter optimisation of the classifiers (for example the Gaussian prior in MAXENT) could be incorporated into the search.

Whilst the relation extractor was successful on intra-sentential relations, it is less successful on inter-sentential relations, perhaps becuase of the lingusitic complexity of these, and the sparsity of positive examples. The split into inter- and inter-sentential examples in the current system seems justified as they have quite different characteristic, but there may also be a case for splitting the intra-sententials further, into intra- and inter-clausals, as suggested by Maslennikov and Chua (2007), and then treating inter-clausals and inter-sententials together. Whilst intra-clausals are more likely to use simple constructions and be amenable to modelling with shallow linguistic features, inter-sententials and inter-clausals are more likely to use complex linguistic phenomena such as corefereces.

Acknowledgements

This work was supported by the Text Mining Programme of ITI Life Sciences Scotland (`http://www.itilifesciences.com`).

References

Bea Alex, Claire Grover, Barry Haddow, Mijail Kabadjov, Ewan Klein, Michael Matthews, Stuart Roebuck, Richard Tobin, and Xinglong Wang. 2008. The ITI TXM Corpora: Tissue Expressions and Protein-Protein Interactions. In *Proceedings of LREC*.

Razvan C. Bunescu and Raymond J. Mooney. 2007. Extracting relations from text: From word sequences to dependency paths. In Anne Kao and Steve Poteet, editors, *Text Mining and Natural Language Processing*, pages 29–44. Springer.

Razvan Bunescu, Ruifang Ge, Rohit J. Kate, Edward M. Marcotte, Raymond J. Mooney, Arun K. Ramani, and Yuk W. Wong. 2005. Comparative experiments on learning information extractors for proteins and their interactions. *Artificial Intelligence in Medicine*, 33(2):139–155.

James Curran and Stephen Clark. 2003. Language independent NER using a maximum entropy tagger. In *Proceedings of CoNLL*.

James Cussens and Claire Nédellec, editors. 2005. *Proceedings of Language Learning in Logic*.

Thomas G. Dietterich. 2000. Ensemble methods in machine learning. *Lecture Notes in Computer Science*, 1857:1–15.

Gunes Erkan, Arzucan Ozgur, and Dragomir R. Radev. 2007. Semi-supervised classification for extracting protein interaction sentences using dependency parsing. In *Proceedings of EMNLP-CoNLL*.

Kuzman Ganchev, Koby Crammer, Fernando Pereira, Gideon Mann, Kedar Bellare, Andrew McCallum, Steven Carroll, Yang Jin, and Peter White. 2007. Penn/UMass/CHOP Biocreative II systems. In *Proceedings of the Second BioCreative Challenge Evaluation Workshop*.

Claudio Giuliano, Alberto Lavelli, and Lorenza Romano. 2006. Exploiting shallow linguistic information for relation extraction from biomedical literature. In *Proceedings of EACL*.

Claire Grover and Richard Tobin. 2006. Rule-based chunking and reusability. In *Proceedings of LREC*.

Isabelle Guyon and André Elisseeff. 2003. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3(Mar):1157–1182.

Jing Jiang and Chengxiang Zhai. 2007. A systematic exploration of the feature space for relation extraction. In *Proceedings of NAACL*.

Thorsten Joachims. 1999. Making large-scale support vector machine learning practical. In *Advances in Kernel Methods: Support Vector Machines*. MIT Press, Cambridge, MA.

S. Katrenko and P. W. Adriaans. 2006. Learning relations from biomedical corpora using dependency tree levels. In *Proceedings of Benelearn*.

Jin D. Kim, Tomoko Ohta, and Jun'ichi Tsujii. 2008. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9(1).

Martin Krallinger, Florian Leitner, Carlos Rodriguez-Penagos, and Alfonso Valencia. 2008. Overview of the protein-protein interaction annotation extraction task of BioCreative II. *Genome Biology (in press)*.

Mstislav Maslennikov and Tat S. Chua. 2007. A multi-resolution framework for information extraction from free text. In *Proceedings of ACL*.

Guido Minnen, John Carroll, and Darren Pearce. 2000. Robust, applied morphological generation. In *Proceedings of INLG*.

Leif Arda Nielsen. 2006. Extracting protein-protein interactions using simple contextual features. In *Proceedings of BioNLP*.

Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Bjorne, Jorma Boberg, Jouni Jarvinen, and Tapio Salakoski. 2007. Bioinfer: A corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8(1).

Marios Skounakis, Mark Craven, and Soumya Ray. 2003. Hierarchical hidden markov models for information extraction. In Georg Gottlob, Toby Walsh, Georg Gottlob, and Toby Walsh, editors, *Proceedings of IJCAI*.

L. Smith, T. Rindflesch, and W. J. Wilbur. 2004. MedPost: a part-of-speech tagger for biomedical text. *Bioinformatics*, 20(14):2320–2321.

Guodong Zhou, Min Zhang, Donghong Ji, and Qiaoming Zhu. 2007. Tree kernel-based relation extraction with context-sensitive structured parse tree information. In *Proceedings of EMNLP-CoNLL*.