Generic Relation Identification: Models and Evaluation

Ben Hachey
Centre for Language Technology
Macquarie University
NSW 2109 Australia

Capital Markets CRC Limited
GPO Box 970
Sydney NSW 2001

bhachey@cmcrc.com

Abstract

Generic relation identification (GRI) aims to build models of relation-forming entity pairs that can be transferred across domains without modification of model parameters. GRI has high utility in terms of cheap components for applications like summarisation, automated data exploration and initialisation of bootstrapping of relation extraction. A detailed evaluation of GRI is presented for the first time, including explicit tests of portability between newswire and biomedical domains. Experimental results show that a novel approach incorporating dependency parsing is better in terms of recall. And, accuracy is shown to be comparable across domains.

1 Introduction

Relation extraction (RE) aims to identify mentions of relations in text. A relation mention is defined as a predicate ranging over two arguments, where an argument represents concepts, objects or people in the world and the predicate describes the type of stative association or interaction that holds between the things represented by the arguments. Input to the RE task consists of source documents with entity mention markup (e.g., Figure 1). The output is a list of relation-forming entity mention pairs and a label indicating the type of relation that holds between them (e.g., Table 1). This paper addresses the relation identification task, which identifies pairs of relation-forming entity mentions (e.g., "David Murray" and "Amidu Berry" in the example).

[^{place} American] saxophonist [^{person} David Murray] recruited [^{person} Amidu Berry] and DJ [^{person} Awadi] from [^{organisation} PBS].

Figure 1: Example input to GRI task (from ACE 2004). Square brackets indicate the extent of entity mentions with type as italicised superscript.

Entity 1	Entity 2	Relation Type
American	David Murray	CITIZEN/RESIDENT
David Murray	Amidu Berry	BUSINESS
David Murray	Awadi	BUSINESS
Amidu Berry	PBS	MEMBER-OF-GROUP
Awadi	PBS	MEMBER-OF-GROUP

Table 1: Example output from GRI task. Relation types are not part of the relation identification task but are given here for purposes of illustration.

Relation extraction (RE) can be addressed using supervised (Zelenko et al., 2005; Blitzer et al., 2006), bootstrapping (Brin, 1998; Riloff and Jones, 1999; Agichtein and Gravano, 2000; Hassan et al., 2006) or generic approaches (Conrad and Utt, 1994; Hasegawa et al., 2004). One way to characterise these different approaches is in terms of adaptation cost, i.e. the amount of work necessary to adapt them to a new domain or task. In these terms, supervised approaches (including rule engineering and supervised machine learning) incur the highest cost as systems need to be built largely from scratch for each new domain. Bootstrapping approaches incur less cost as they require only a small amount of seed data. And generic approaches provide domain adaptation for free as parameters do not need to be modified for new domains or tasks. Another way to char-

acterise these approaches is in terms of the ontology creation problems they address, i.e. whether they address only the instantiation task where instances are added to an ontology in a new domain given a *relation schema* (the taxonomy of relation types to be identified) or whether they also address the task of learning the relation schema for the new domain. In these terms, supervised approaches and bootstrapping approaches address only the ontology instantiation problem while generic approaches also address the problem of learning relation schemas from data. The trade-off is in terms of accuracy, where generic approaches suffer when compared to supervised and bootstrapping approaches. However, generic approaches have high utility in terms of developing cheap components for applications like paraphrase acquisition (Hasegawa et al., 2005), on-demand information extraction (Sekine, 2006) and automatic summarisation (Hachey, 2009a). And, they could be used for initialisation of semi-supervised bootstrapping of relation extraction.

This paper contains the first detailed evaluation of generic relation identification (GRI), including explicit tests of portability between newswire and biomedical domains. GRI can be split into two sub-tasks, where input consists of source documents with entity mention markup (as in Figure 1). The first sub-task has the goal of identifying relation-forming entity mention pairs and outputs a list of co-occurring entity mention pairs (e.g., Table 1). The second sub-task has the goal of applying a ranking over co-occurring pairs that indicates the strength of association. This ranking might be used for filtering low confidence relations or in weighting schemes for extrinsic applications (e.g., automatic summarisation). The experiments here focus primarily on the identification sub-task, which is evaluated with respect to gold standard data. Experiments are reported that compare window-based models (e.g., setting a threshold on the number of intervening tokens). Results show that a novel approach incorporating intervening words and dependency paths is better in terms of recall while being statistically indistinguishable in terms of precision and f-score. Furthermore, performance is shown to be comparable when porting from news to biomedical text without modification of model parameters.

Author	Co-occur Window	Constraints
Hasegawa	W/in 5 words	NA
Zhang	Sentence	Spanning parse
Conrad	W/in 25, 100 words	NA
Smith	Sentence	NA
Filatova	Sentence	Verbal connector

Table 2: Approaches from the GRI literature.

2 Related Work

Table 2 contains an overview of approaches from the GRI literature. The first column (Author) contains the first author of the approaches referenced in the following text. The first two rows correspond to approaches that address relation identification and characterisation; the third and fourth rows correspond to approaches that focus on the GRI task; and the fifth row corresponds to a related approach to generic *event* identification and characterisation. The second column (Co-occur Window) describes the co-occurrence window for identifying entity mention pairs (e.g., W/in 5 words means that entity mention pairs need to occur within five tokens of each other). The third column (Constraints) describes any additional constraints placed on entity mention pairs.

Hasegawa et al. (2004) introduce the task of relation discovery (using unsupervised techniques to annotate pairs of associated objects with a relation type derived from the textual context). Their work includes a simple approach to GRI where all pairs of entity mentions within 5 tokens of each other are considered to be co-occurring. No motivation is given for choosing 5 as the threshold. In subsequent work, Zhang et al. (2005) incorporate syntactic parsing (Collins, 1999) into their approach to GRI. All pairs of entities in the same sentence are considered to be co-occurring provided that there is a spanning parse. Neither Hasegawa et al. nor Zhang et al. explicitly evaluate their approaches to relation identification.

Filatova and Hatzivassiloglou (2003) describe related work that aims to extract entity pair associations that constitute what they term atomic events. They consider any pair of entity mentions co-occurring within a sentence to be possible participants in event descriptions and they add a constraint requiring that a verbal 'connector' (i.e., a verb or a noun that is a WordNet hy-

ponym of *event* or *activity*) be present in the intervening token context between the entity mentions. The authors present a limited evaluation of their approach to relation identification which suggests reasonable precision. However, it is based on manual analysis of the system output so is not repeatable. Furthermore, it does not address recall and it does not compare the system to any lower or upper bounds on accuracy.

Conrad and Utt (1994) present seminal work on mining pairs of entities from large text collections using statistical measures of association to rank named entity pairs based on co-occurrence. They propose windows of size 25 and 100, which means that any other entity mention within 25 or 100 tokens to the right or left of a given entity mention is considered to co-occur. These window sizes are chosen as they roughly approximate mean sizes of paragraphs and documents in their data. The authors do not specify which window size they use for their evaluation. A manual evaluation of system output is reported, which suggests reasonable performance but is not repeatable.

Smith (2002) considers all pairs of entities in the same sentence to be co-occurring. He performs an evaluation using a corpus of nineteenth century American historical documents. Extracted entity pairs are compared to a curated resource, which contains expert assessments of the severity of battles in the American civil war. Again, this suggests reasonable performance but is not repeatable. Furthermore, Smith (2002) does not compare to lower or upper bounds.

In the literature on supervised relation extraction, e.g. Liu et al. (2007), features based on parse trees have been used successfully. However, beyond requiring a spanning parse tree (Zhang et al., 2005), no previous approaches have investigated the use of syntactic parsing to constrain GRI. The current work investigates the use of domain-neutral co-occurrence windows for GRI that are based on paths connecting entity mention pairs through syntactic parse trees. Furthermore, it presents the first detailed evaluation of GRI on publicly available relation extraction data.

3 Evaluation

To address previous shortcomings, a principled framework is introduced that uses gold standard relation extraction data to optimise and evaluate GRI models. This is derived from news data from the Automatic Content Extraction (ACE) 2004 and 2005 shared tasks[1] and biomedical data derived from the BioInfer corpus.[2] The ACE 2004 data is used for development experiments. The ACE 2005 data serves as the held-out news test set and the BioInfer data serves as the biomedical test set. See Hachey (2009b) for details of the data preparation and experimental setup.

GRI T/F	ACE2004	ACE2005	BioInfer
True	949	558	1591
False	8304	5587	4252
Total	9253	6145	5843

Table 3: Distribution of relations.

Accuracy is measured in terms of precision (P) and recall (R):

$$P = \frac{NumCorrect}{TotalSystemPairs} \qquad R = \frac{NumCorrect}{TotalGoldPairs}$$

And, f-score (F) is calculated in the standard way: $F = 2PR/(P + R)$. Paired Wilcoxon signed ranks tests[3] across entity pair sub-domains are used to check for significant differences between systems. Sub-domains are formed by taking just those relations between two entities of given types (e.g., *Person-Organisation*, *Gene-Protein*). Table 3 contains the count of same-sentence entity mention pairs that constitute relation mentions (True) and those that are not (False). In the ACE 2004 and 2005 data sets, this results respectively in 949 and 558 true relation mentions spread across seven entity pair sub-domains. In the BioInfer data set, this results in 1591 true relation mentions spread across seven entity pair sub-domains.

The evaluation here also introduces an upper bound for GRI based on human agreement. This is calculated by first obtaining a mapping from entity mentions marked by annotators to entity mentions in the adjudicated gold standard annotation. The mapping used here is derived from the ACE 2005 evaluation script, which computes an

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

[2] `http://mars.cs.utu.fi/BinInfer`

[3] The paired Wilcoxon signed ranks test is a non-parametric analogue of the paired t test. The null hypothesis is that the two populations from which the scores are sampled are identical. Following convention, the null hypothesis is rejected for values of p less than or equal 0.05.

optimised one-to-one mapping based on maximal character overlap between system and gold standard entity mention strings. Given this mapping, it is possible to determine for each possible entity mention pair whether the annotators marked a relation mention. Interestingly, the annotators have high agreement with the adjudicated data set in terms of precision at 0.906 and lower agreement in terms of recall at 0.675. This suggests that the annotators rarely marked bad relation mentions but each missed a number of relation mentions that the other annotator marked. The mean human f-score agreement is 0.773.

4 Models

The GRI task can be generalised in terms of the GENERICRELATIONID algorithm in Figure 2. This takes as input an array of entity mentions E and the Boolean function ISPAIR. The ISPAIR function returns true if two entity mention indices constitute a co-occurring pair and false otherwise. Figure 2 includes the ISPAIR$_{baseline}$ function as an example, which simply counts all pairs of entity mentions occurring in the same sentence as relation-forming pairs like Smith (2002). The GENERICRELATIONID algorithm starts by initialising the set of entity mention pairs $\mathcal{P}$ to the empty set. It then loops over all possible pairs of entities from E, which is assumed to be sorted in terms of the order of occurrence. Pairs are added to $\mathcal{P}$ if the text describes a relation between them. The experiments here will be based on different definitions of the ISPAIR function, based on intervening token windows and dependency path windows.[4]

Atomic Events The first model of entity mention co-occurrence is based on the approach to identifying atomic events from Filatova and Hatzivassiloglou (2003). This uses an ISPAIR$_{event}$ function that accepts all pairs of entity mentions that 1) occur in the same sentence and 2) have a verbal 'connector' (i.e., a verb or a noun that is a WordNet hyponym of *event* or *activity*) in the intervening context.

[4]Additional experiments not reported here also explored learnt ISPAIR functions using decision trees and various combinations of generic features. However, these models did not generalise across domains.

```
GENERICRELATIONID: E, ISPAIR
1   P ← {}
2   i ← 0
3   while i ≤ length(E)
4         j ← i + 1
5         while j ≤ length(E)
6               if ISPAIR(i, j)
7                     P ← P ∪ [i, j]
8               j ← j + 1
9         i ← i + 1
10  return P
```

```
ISPAIR_baseline : i, j
1   if sent(i) = sent(j)
2         return true
3   else
4         return false
```

Figure 2: Algorithm for generic relation identification with baseline function for identifying co-occurring entity mention pairs.

Intervening Token Windows The next model is based on intervening token windows (Toks). It uses an ISPAIR$_{toks}$ function that counts all pairs of entity mentions that 1) occur in the same sentence and 2) have t or fewer intervening tokens. Most previous GRI work has used some variant of this model. Hasegawa et al. (2004), for example, use the ISPAIR$_{toks}$ function but do not motivate their threshold of t=5.

Figure 3 contains optimisation results for setting the intervening token threshold t on the news development data (ACE 2004). The shaded bars correspond to mean f-scores (actual value printed above the bars) for different settings of t (specified along the bottom of the horizontal axis). The best f-score is shown in bold. Values that are statistically distinguishable from the best ($p \leq 0.05$) are underlined. The results suggest that the best performance is achieved with t set to 2, though this is not reliably different from scores for t=1 and t=4 which suggests a range of optimal values from 1 to 4. For the comparisons in the rest of this paper, the Toks model should be assumed to have t set to 2 unless stated otherwise. Recall (R) and precision (P) are plotted as dotted grey and solid black lines respectively and are closest to being balanced at t=1.

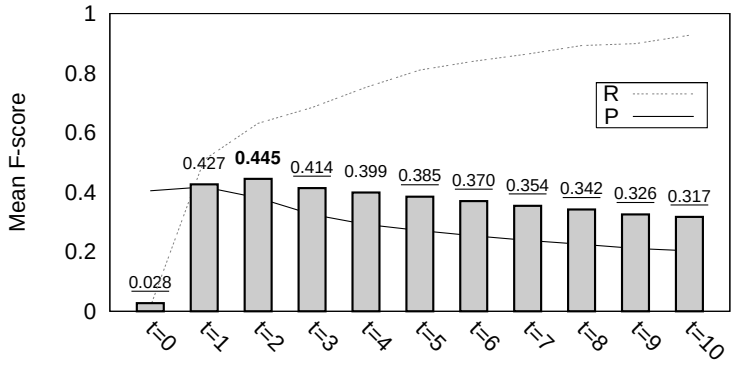

Figure 3: Window size results for token-based model.

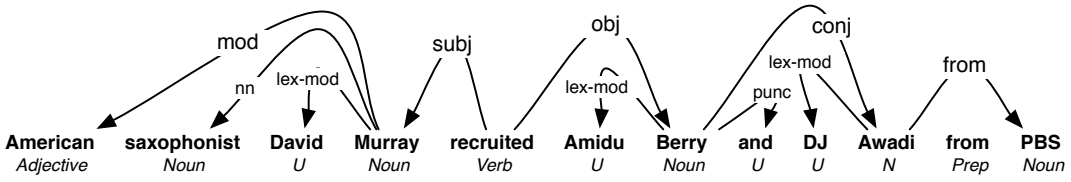

Figure 4: Dependency parse for example sentence.

Dependency Path Windows The experiments here also consider a novel approach to modelling entity mention co-occurrence that is based on syntactic governor-dependency relations (Deps). This uses an IsPair$_{deps}$ function that counts all pairs of entity mentions that 1) occur in the same sentence and 2) have d or fewer intervening token nodes on the shortest dependency path connecting the two entity mentions. Dependency paths are derived using the Minipar software (Lin, 1998), which produces 1) directional links from governors to their dependent lexical items and 2) grammatical relation types (e.g., *subject*, *object*). Figure 4 contains the Minipar parse of the example sentence from Figure 1. The shortest dependency paths between all candidate entity mention pairs are extracted from the parse graph. The path between "American" and "David Murray", for example, consists of a direct *modifier* (mod) relation with zero intervening word token nodes. The path between "David Murray" and "Awadi", on the other hand, passes through one word token node ("recruited") after post-processing operations that pass governor-dependency relations along chains of conjoined tokens, resulting in a *obj* relation between recruited and Awadi.

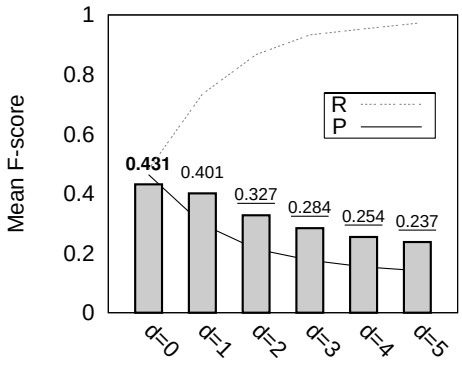

Figure 5: Window size results for dependency-based model.

Figure 5 contains optimisation results for setting the dependency path threshold d on the news development data (ACE 2004). The shaded bars correspond to mean f-score. The best f-score is shown in bold and is achieved at d=0 (which should be assumed from here). Values that are statistically distinguishable are underlined. Results here suggest a range of optimal values from d=0 to d=1. Recall (R) and precision (P) are plotted as dotted grey and solid black lines respectively and are closest to being balanced at d=0.

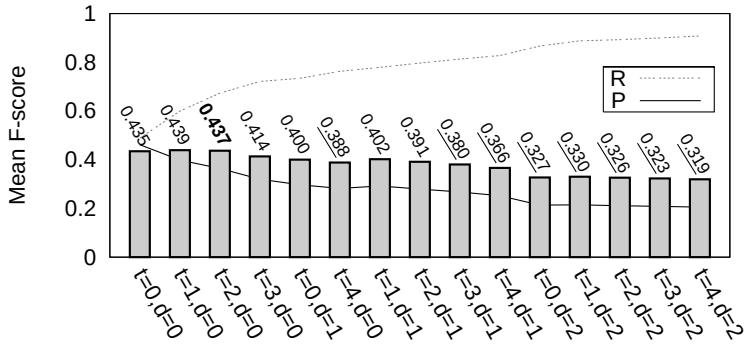

Figure 6: Window size results for combined (token and dependency) model.

Combined Windows Finally, the current work also introduces an entity mention co-occurrence model that combines token and dependency windows (Comb). It uses an IsPair$_{comb}$ function that counts all pairs of entity mentions that 1) occur in the same sentence and 2) either have t or fewer intervening tokens or have d or fewer intervening dependency path nodes. Based on tuning experiments on the news development data (ACE 2004), the thresholds here are set to t=1 and d=0.

Figure 6 contains joint optimisation results for the intervening token (t) and dependency path (d) thresholds on the news development data (ACE 2004). The optimal system is chosen in terms of the mean rank of f-scores across entity pair subdomains. The best mean rank is achieved with t=2 and d=0. Values that are statistically distinguishable from the best are underlined. The results suggest a range of optimal settings with t ranging from 0 to 2 and d ranging from 0 to 1.

5 Results

Table 4 contains P, R and F results. The best score for each measure is in bold and scores that are statistically distinguishable from the best ($p \leq 0.05$) are underlined. The baseline system considers all pairs in the same sentence to be relations.

Which window function is best for identifying relation mentions? The highest f-score on the news test data is obtained using the dependency path model, though this is not statistically distinguishable from the Toks or Comb models. In terms of recall, the Comb model obtains the highest score (0.538), which is significantly better than the Toks and Deps models. The Deps model, however, obtains a precision score that is significantly better than the Comb model. For the current work, the combined model is considered to be the best as it achieves the highest recall while the f-score is statistically indistinguishable from the other models. The prioritisation of recall is motivated by the fact that weighting is generally applied to co-occurring entity pairs for applications of GRI. For example, relation mining approaches from the related literature (Conrad and Utt, 1994; Smith, 2002) use statistical measures of association such as pointwise mutual information, ϕ^2 and log likelihood ratio to estimate association strengths. Thus, a certain amount of noise in GRI should be acceptable if the subsequent weighting scheme is assumed to give higher weight to true relation-forming entity pairs.

How does system performance compare to human performance? The main difference is in terms of precision, where the Comb model performs far worse than the Human upper bound (0.906). However, while Comb recall is significantly worse than Human recall (0.675), the difference is not large. Furthermore, inter-annotator agreement on ACE is a very strong upper bound for the GRI task as the annotators are given detailed guidelines that provide a prescriptive notion of what counts as a relation mention. The GRI task, on the other hand, is not guided by a predefined schema and GRI predicts a number of relation mentions that are incorrect with respect to the gold standard annotation but could arguably be considered true relation mentions.

a) ACE 2005 (News Test Set)

	P	R	F
Baseline	0.110	1.000	0.195
Event	0.050	0.392	0.083
Toks	0.291	0.510	0.342
Deps	**0.456**	0.392	**0.360**
Comb	0.277	**0.538**	0.332
Human	0.906	0.675	0.773

b) BioInfer (Biomedical Test Set)

	P	R	F
Baseline	0.268	1.000	0.415
Event	0.186	0.418	0.247
Toks	**0.527**	0.388	0.422
Deps	0.450	0.302	0.349
Comb	0.500	**0.454**	**0.453**
Human	NA	NA	NA

Table 4: Comparison of P, R and F on news and biomedical test sets. The best score in each column is in bold and those that are statistically distinguishable from the best are underlined.

Does model performance generalise across domains? In the biomedical domain, the Comb model performs best in terms of f-score with a score of 0.453 though it is statistically indistinguishable from the Toks model. This is a stronger result than in the news domain where there was no significant differences among the f-scores of the Toks, Deps and Comb models. Consistent with the news domain, there are no significant differences among the precision scores of the Toks, Deps and Comb models and, importantly, the Comb model is significantly better than the Toks and Deps models in terms of recall in both domains. Interestingly, the f-score of the Baseline model is statistically indistinguishable from the Comb model on the biomedical data. Since Baseline recall is the same for both domains (1.000), this is due to higher precision (0.268 as opposed to 0.110). This suggests that the biomedical GRI task is easier due to the higher proportion of true relation-forming pairs (27% compared to approximately 10% for the ACE data sets). This may be artificially high, however, since the BioInfer creators selectively sampled sentences that include mentions of proteins that are known to interact. The biomedical result is consistent with the news result, however, in that Comb precision is significantly better than Baseline precision on both domains.

6 Discussion

Recall and precision of the Event model The low recall of the Event model with respect to the other models is not surprising due to the constraint requiring an intervening event word. The low precision, however, indicates that the constraint is not helpful as a method to capture long-distance relation mentions based on intervening token windows. The Event model does particularly poorly on the ACE 2005 *GPE-GPE* and BioInfer *Protein-ProteinFamily* entity pair sub-domains due to the fact that true pairs rarely have a verbal connector in the intervening token context. True relation mentions in the ACE 2005 *GPE-GPE* sub-domain tend to be geographical part-of relations where the two entity mentions are adjacent (e.g., the relation between the *GPE* entity mention "Peoria" and the *GPE* entity mention "Illinois" in the fragment "Peoria, Illinois"). And, true relation mentions in the BioInfer *Protein-ProteinFamily* sub-domain tend to be appositives (e.g., the relation between the *Protein* entity mention "cofilin" and the *ProteinFamily* entity mention "actin-binding protein" in the fragment "cofilin, a ubiquitous actin-binding protein") or nominal modifiers (e.g., the relation between the *ProteinFamily* entity mention "cyclin-dependent kinase inhibitors" and the *Protein* entity mention "p57" in the fragment "the cyclin-dependent kinase inhibitors (CKIs) p27 and p57").

Error Analysis For each entity pair sub-domain, ten instances were chosen randomly from the set of erroneously classified instances. These were manually inspected in order to characterise the types of errors made by the combined (Comb) GRI system. This suggests that the majority of false positive errors in both the news and biomedical data sets (81% and 54% respectively) can be considered implicit relation mentions (i.e., the relation is not explicitly stated but

is more or less implicit given the context of the sentence). For example, our system posits a false positive relation between "Gul" and "Erdogan" in the sentence "Unlike the soft-spoken Gul, Erdogan has a reputation as a fighter." These types of false positives are not necessarily problematic in applications of GRE. In fact, these implicit relation mentions are likely to be helpful in applications, e.g. representing the conceptual content of a sentence for extractive summarisation (Hachey, 2009a). One not unexpected difference between domains is that there were considerably more false negative errors in the biomedical data that could be attributed to parsing errors (15% as opposed to 5% in the news data).

Comparison of ranking methods Since it is trivial to improve recall simply by increasing token or dependency thresholds, improvements in f-scores require models with higher precision. One possible approach for improving precision would be to incorporate methods from the literature (Conrad and Utt, 1994; Smith, 2002) for ranking entity mention pairs using statistical measures of association, such as pair probability (Pr), log-likelihood (G^2), ϕ^2, and pointwise mutual information (PMI). Table 5 contains correlation (point-biserial) scores that compare rank weights obtained from these measures with a binary variable indicating whether the instance constitutes a true relation mention according to the annotation. Following Cohen (1988), values over 0.10 (typeset in italicised bold font) are considered to indicate a small effect and values over 0.30 (typeset in bold font) are considered to indicate a medium effect. The table suggests that a threshold filtering low values of PMI would be the best filter for the ACE 2005 test set (small to medium correlation of 0.273, 0.356, 0.168 and 0.326 respectively for the Baseline, Toks, Deps and Comb models). On the BioInfer test set, by contrast, no measure has consistent correlation across systems and effect sizes are largely negligible. The highest correlation is 0.116 for G^2 on the Comb system. While this effect is small, in conjunction with the ACE 2005 results, it suggests that G^2 would be the better ranking method for domain-neutral relation identification.

a) ACE 2005 (News Test Set)

	Pr	G^2	ϕ^2	PMI
Baseline	-0.093	***0.108***	***0.262***	***0.273***
Toks	-0.098	***0.250***	**0.329**	**0.356**
Deps	-0.092	0.067	***0.145***	***0.168***
Comb	-0.091	***0.219***	***0.294***	**0.326**

b) BioInfer (Biomedical Test Set)

	Pr	G^2	ϕ^2	PMI
Baseline	0.030	0.037	***0.105***	0.073
Toks	***0.114***	***0.107***	-0.009	-0.004
Deps	0.056	0.070	-0.023	-0.008
Comb	0.081	***0.116***	0.003	0.041

Table 5: Point-biserial correlation analysis comparing a true relation mention indicator feature to various approaches for ranking GRI predictions by pair association strength.

7 Conclusions

This paper presented a detailed evaluation of the generic relation identification (GRI) task, providing a comparison of various window-based models for the first time. It compared the intervening token window approach (Toks) from the literature to a novel GRI approach based on windows defined over dependency paths (Deps). In addition, it introduced a combined approach (Comb) that integrates the intervening token and dependency path models. Models were optimised on gold standard data in the news domain and applied directly to data from the news and biomedical domains for testing. The use of the ACE 2005 data for a news test set allowed comparison to a human upper bound for the task.

Model comparison suggested that the Deps and Comb models are best. In particular, the Comb approach performed reliably better than the other models in terms of recall while maintaining statistically indistinguishable precision and f-score. High recall models were prioritised here based on the fact that applications of generic relation extraction generally incorporate a mechanism for ranking identified relation mentions. Experiments and analysis suggest that GRI accuracy is comparable when applying the newswire-optimised models directly to the biomedical domain.

Acknowledgments

This work was supported by Scottish Enterprise Edinburgh-Stanford Link grant R37588 as part of the EASIE project at the University of Edinburgh. It would not have been possible without the guidance of Claire Grover and Mirella Lapata. I would also like to thank Robert Gaizauskas and Steve Renals for thoughtful feedback.

References

Eugene Agichtein and Luis Gravano. 2000. Snowball: Extracting relations from large plain-text collections. In *Proceedings of the 5th ACM International Conference on Digital Libraries*, San Antonio, TX, USA.

John Blitzer, Mark Dredze, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia.

Sergey Brin. 1998. Extracting patterns and relations from the world wide web. In *Proceedings of the EDBT International Workshop on the Web and Databases*, Valencia, Spain.

Jinxiu Chen, Donghong Ji, Chew Lim Tan, and Zhengyu Niu. 2005. Automatic relation extraction with model order selection and discriminative label identification. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing*, Jeju Island, Korea.

Jinxiu Chen, Donghong Ji, Chew Lim Tan, and Zhengyu Niu. 2006. Unsupervised relation disambiguation with order identification capabilities. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia.

Jacob Cohen. 1988. *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates, Inc., San Diego, CA, second edition.

Michael Collins. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania.

Jack G. Conrad and Mary Hunter Utt. 1994. A system for discovering relationships by feature extraction from text databases. In *Proceedings of the 17th SIGIR*, Melbourne, Australia.

Elena Filatova and Vasileios Hatzivassiloglou. 2003. Marking atomic events in sets of related texts. In Nicolas Nicolov, Kalina Bontcheva, Galia Angelova, and Ruslan Mitkov, editors, *Recent Advances in Natural Language Processing III*. John Benjamins, Amsterdam/Philadelphia.

Ben Hachey. 2009a. Multi-document summarisation using generic relation extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Singapore.

Ben Hachey. 2009b. *Towards Generic Relation Extraction*. Ph.D. thesis, University of Edinburgh.

Takaaki Hasegawa, Satoshi Sekine, and Ralph Grishman. 2004. Discovering relations among named entities from large corpora. In *Proceedings of the 42nd ACL*, Barcelona, Spain.

Takaaki Hasegawa, Satoshi Sekine, and Ralph Grishman. 2005. Unsupervised paraphrase acquisition via relation discovery. Technical report, Proteus Project, Computer Science Department, New York University.

Hany Hassan, Ahmed Hassan, and Sara Noeman. 2006. Graph based semi-supervised approach for information extraction. In *Proceedings of the TextGraphs: The 2nd Workshop on Graph Based Methods for Natural Language Processing*, New York, NY, USA.

Tor-Kristian Jenssen, Astrid Lægreid, Jan Komorowski, and Eivind Hovig. 2001. A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*, 28(1):21–28.

Dekang Lin. 1998. Dependency-based evaluation of MINIPAR. In *Proceedings of the LREC Workshop Evaluation of Parsing Systems*, Granada, Spain.

Yudong Liu, Zhongmin Shi, and Anoop Sarkar. 2007. Exploiting rich syntactic information for relationship extraction from biomedical articles. In *Proceedings of NAACL-HLT*, Rochester, NY, USA.

Ellen Riloff and Rosie Jones. 1999. Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the 16th National Conference on Artificial Intelligence*, Orlando, FL, USA.

Satoshi Sekine. 2006. On-demand information extraction. In *Proceedings of the COLING/ACL Main Conference Poster Sessions*, Sydney, Australia.

David A. Smith. 2002. Detecting and browsing events in unstructured text. In *Proceedings of the 25th SIGIR*, Tampere, Finland.

Dmitry Zelenko, Chinatsu Aone, and Jason Tibbetts. 2005. Trainable evidence extraction system (TEES). In *Internationl Conference on Intelligence Analysis*, McLean, VA, USA.

Min Zhang, Jian Su, Danmei Wang, Guodong Zhou, and Chew Lim Tan. 2005. Discovering relations from a large raw corpus using tree similarity-based clustering. In *Proceedings of the 2nd IJCNLP*, Jeju Island, Korea.