Towards Generic Relation Extraction

Benjamin Hachey

Doctor of Philosophy

Institute for Communicating and Collaborative Systems

School of Informatics

University of Edinburgh

2009

Abstract

A vast amount of usable electronic data is in the form of unstructured text. The relation extraction task aims to identify useful information in text (e.g., PersonW works for OrganisationX, GeneY encodes ProteinZ) and recode it in a format such as a relational database that can be more effectively used for querying and automated reasoning. However, adapting conventional relation extraction systems to new domains or tasks requires significant effort from annotators and developers. Furthermore, previous adaptation approaches based on bootstrapping start from example instances of the target relations, thus requiring that the correct relation type schema be known in advance. Generic relation extraction (GRE) addresses the adaptation problem by applying generic techniques that achieve comparable accuracy when transferred, without modification of model parameters, across domains and tasks.

Previous work on GRE has relied extensively on various lexical and shallow syntactic indicators. I present new state-of-the-art models for GRE that incorporate governor-dependency information. I also introduce a dimensionality reduction step into the GRE relation characterisation sub-task, which serves to capture latent semantic information and leads to significant improvements over an unreduced model. Comparison of dimensionality reduction techniques suggests that latent Dirichlet allocation (LDA) – a probabilistic generative approach – successfully incorporates a larger and more interdependent feature set than a model based on singular value decomposition (SVD) and performs as well as or better than SVD on all experimental settings. Finally, I will introduce multi-document summarisation as an extrinsic test bed for GRE and present results which demonstrate that the relative performance of GRE models is consistent across tasks and that the GRE-based representation leads to significant improvements over a standard baseline from the literature.

Taken together, the experimental results 1) show that GRE can be improved using dependency parsing and dimensionality reduction, 2) demonstrate the utility of GRE for the content selection step of extractive summarisation and 3) validate the GRE claim of modification-free adaptation for the first time with respect to both domain and task. This thesis also introduces data sets derived from publicly available corpora for the purpose of rigorous intrinsic evaluation in the news and biomedical domains.

Acknowledgements

First, I would like to thank my supervisors. I would not have done this PhD if it had not been for the encouragement, support, advice and feedback I received from Claire Grover. Likewise, the advice and feedback I received from Mirella Lapata was invaluable. I would also like to thank Robert Gaizauskas and Steve Renals for a very thoughtful and rewarding thesis defence.

A number of other people provided feedback on various documents and half-baked thoughts during my PhD. In particular, I would like to thank Frank Keller and Simon King for their considerable contribution to my draft dissertation defence committee. And I would like to thank Sam Brody, Trevor Cohn, Sebastian Riedel and Simone Teufel for taking the time to discuss aspects of this work.

I was incredibly lucky to meet a number of other people in Edinburgh who I consider to be friends and mentors. I would like to thank Amy Isard and Colin Matheson for being great office mates and for helping to make sure I ate on Tuesdays and drank on Fridays. I was lucky to work with Ewan Klein, Jon Oberlander and Henry Thompson. And I was very lucky to be able to collaborate with Beatrice Alex, Markus Becker, Gabriel Murray and David Reitter.

I would like to thank Elena Filatova, Satoshi Sekine and their collaborators for providing the starting point for this work and for taking the time to correspond concerning the details of their own work. I am also grateful to the people responsible for creating and sharing the corpora used in this work. And I am grateful to Richard Tobin for developing and sharing the LT XML tools.

I gratefully acknowledge the financial support provided by the Edinburgh-Stanford Link programme. I would also like to thank Chris Manning for hosting me when I visited Stanford and for discussing early ideas that went into this thesis.

Finally, I would not have been in a position to do this work if it had not been for the support and encouragement of my friends and family. I would like to thank my siblings and especially my parents, who passed on to me their enthusiasm for learning. More than anything, I can not thank Anna enough for her grace and good humour.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Benjamin Hachey*)

Preface

Parts of the work reported in Chapters 3 and 5 of this thesis appear in published proceedings (Hachey, 2006, 2007; Hachey et al., 2008).

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

> "In Ersilia, to establish the relationships that sustain the city's life, the inhabitants stretch strings from the corners of the houses, white or black or gray or black-and-white according to whether they mark a relationship of blood, of trade, authority, agency."
>
> Italo Calvino, *Invisible Cities*

1.1 What is a Relation?

A vast amount of usable electronic data is in the form of unstructured text. The information extraction task aims to identify useful information in text and recode it in a format such as a relational database that can be more effectively used for querying and automated reasoning (e.g., Turmo et al., 2006). Typically, this extraction task includes the sub-tasks of identifying named objects (e.g., persons, organisations, dates), identifying relationships between named objects (e.g., *PersonX* works for *OrganisationY*), and identifying events (e.g., *PersonX* was hired by *OrganisationY* on *DateZ*). The current work addresses the second task which is referred to as relation extraction (RE). The RE task aims to identify mentions of relations in text,[1] where a relation mention is defined as a predicate ranging over two arguments, where an argument represents concepts, objects or people in the real world and the relation predicate describes the type of stative association or interaction that holds between the things represented by the arguments.

Figure 1.1 contains example relation mentions from the news and biomedical data sets used in the experimental chapters of this thesis. The left side of the figure con-

[1] Other modalities such as speech can also be considered, but the work here focuses on text (including newswire, broadcast news transcripts and scientific papers in the biomedical domain).

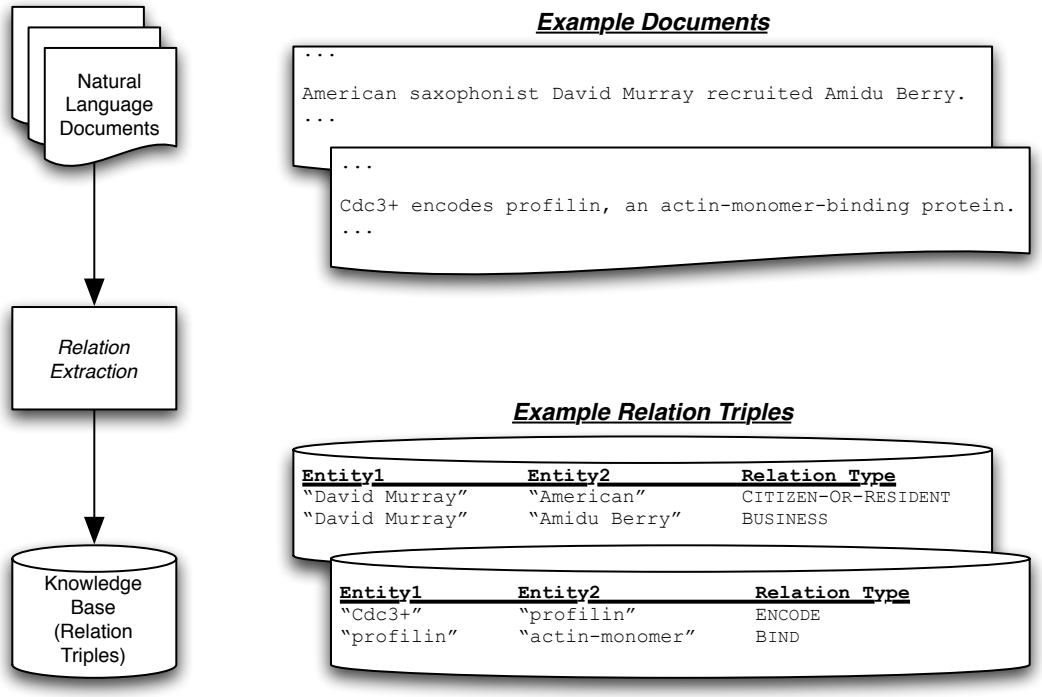


Figure 1.1: *Overview of relation extraction task with example input and output.*

tains a pipeline representation of the RE task. The input consists of natural language documents containing e.g. unstructured text or speech. These documents are fed to the RE system, which identifies and characterises the relations described in the text or speech data. The output of the RE system consists of relation mention triples which include the two entity mentions that take part in the relation and the relation type. The right side of Figure 1.1 contains two example input documents on the top and the relation mention triples from those sentences on the bottom. The first document contains the sentence "American saxophonist David Murray recruited Amidu Berry". This contains two relation mentions: 1) a reference to a CITIZEN-OR-RESIDENT relation between "David Murray" and "American" and 2) a reference to a BUSINESS relation between "David Murray" and "Amidu Berry". Likewise the sentence in the second document contains two relation mentions: 1) a reference to an ENCODE relation between "Cdc3+" and "profilin" and 2) a reference to a BIND relation between "profilin" and "actin-monomer".

1.2 Relation Extraction and Adaptation

A majority of the systems developed to address the relation extraction task are based on either rule engineering or supervised machine learning, both of which are expensive to port to new domains. In the case of rule engineering, writing extraction rules requires extensive effort from a rule engineering expert who is familiar with the target domain. In the case of supervised learning, annotation of training data and tuning features/model parameters require extensive effort from at least one annotator (expert in the target domain) and from a natural language processing expert.

The expense of conventional supervised approaches has motivated another vein of work on bootstrapping and transfer learning. One prominent approach is initialised with only a small seed set of example relation-forming entity pairs for a particular relation type. A wide-coverage system is then bootstrapped through an iterative process of inducing extraction rules given entity pairs and subsequently identifying new entity pairs given extraction rules (e.g., Brin, 1998; Agichtein and Gravano, 2000). Other partially supervised approaches include active learning where a conventional relation extraction system is trained on a small seed corpus, after which it chooses examples that are difficult to classify to be presented to a human annotator (e.g., Zelenko et al., 2005). The expense of conventional supervised approaches has also motivated recent work in other areas of natural language processing on transfer learning (e.g., Blitzer et al., 2006; Daumé III, 2007) and domain adaptation (e.g., Nivre et al., 2007).

Transfer learning and partially supervised approaches, however, still require the relation type schema to be known in advance for each new domain. Conrad and Utt (1994) and Hasegawa et al. (2004) present pioneering studies using generic approaches for two relation extraction sub-tasks: entity association mining and discovery of typed relations. These approaches make the leap from simply instantiating databases (or ontologies) based on predetermined schema to automatically learning the relation type schema for a new domain and provide domain adaptation for free through the use of generic techniques that can be transferred without modification of model parameters (i.e., with no annotation in the new domain). However, while it is a key motivation, previous work on these tasks has largely failed to explicitly evaluate the claim of modification-free adaptation to new domains or tasks. This thesis synthesises the previously disjoint bodies of literature on relation mining and relation discovery into the combined generic relation extraction (GRE) framework, introduces new state-of-the-art approaches and explicitly demonstrates portability across domains and tasks.

1.3 Utility of Generic Relation Extraction

Despite two decades of work on the task, the utility of relation extraction remains largely theoretical. This is not to say that the task is not well motivated but rather that there have only recently been scientific studies explicitly testing the contribution of relation extraction in a controlled environment (e.g., Alex et al., 2008a). Given that a primary motivation for early work was extracting structured information for database curation and information analysis, a first evaluation of the utility of RE should perhaps look at whether it helps human analysts in information gathering tasks.[2] However, this is an expensive undertaking as it involves access to analysts as well as IE systems. Furthermore, there is an intricate interplay between the user interface and the IE technology that makes it difficult to isolate the effect of IE (Karamanis, 2007). For this reason, it is difficult to directly measure the effect of IE on database curation and the community has been slow to publish such studies. However, there are several other applications that could serve as frameworks for evaluating the utility of RE technology.

One way to view the result of information extraction is as a social network, i.e. a graph of relationships that indicate the important entities in a domain and can be used to study or summarise interactions. The extracted social networks could be used to create biographical sketches for entities which can then be exploited for summarisation and question answering (e.g., Jing et al., 2007). The networks could also provide an alternative to standard presentation of information retrieval results when interacting with a document collection, e.g. by providing browsable representation of entities and relationships that link to documents where they are described. For example, Sekine (2006) suggests creating table based summaries of relations in query results. In a similar vein, social networks could be extracted from a document collection offline and then used for search. For example, Agichtein et al. (2005) suggest pre-collecting commonly queried relations for factoid question answering. Similar information could also be used to provide search functionality where a user enters two entities and the social network is used to identify the shortest or most likely paths connecting the two (or more) entities.

Finally, the results of RE can also be used as input to other NLP tasks. For example, Hasegawa et al. (2005) use GRE for paraphrase acquisition, automatically discovering different ways to express the same relation. A related application of RE is as a rep-

[2] Work summarised by Smalheiser and Swanson (1998) does prove the utility of relation extraction, by reporting cases where it is successful in predicting useful relations. However, this falls short of quantifying the accuracy and utility of relation extraction.

resentation for automatic summarisation. Here, relation information can be used to represent the underlying semantics of a document. This representation can be used in conjunction with existing extractive summarisation techniques to identify sentences expressing salient relations that should be part of the summary. In addition to evaluating the portability of relation extraction models with respect to domain, this thesis also explicitly demonstrates the utility of GRE for extractive summarisation, using the relation models developed in this thesis as a conceptual representation for modelling sentence semantics.

1.4 Contributions and Thesis Outline

In Chapter 2, the GRE task is situated within the broader literature on information extraction. First, the relation extraction task is presented in its historical context and defined. Deployment and engineering requirements of various approaches from the literature are discussed which motivate generic and bootstrapping approaches. Next, previous approaches to generic relation extraction are discussed, covering the literatures on named entity association mining and relation discovery. A summary discussion serves to identify several shortcomings of previous evaluations and areas where models can be improved.

In Chapter 3, the task is formalised in a generalised framework that unifies the previously disjoint literatures on named entity association mining and discovery of typed relations. Data sets that are derived from publicly available corpora are described for evaluation in the news and biomedical domains. This allows the GRE claim of modification-free adaptation with respect to domain to be explicitly evaluated for the first time by applying news-optimised models directly to the biomedical domain. Double annotation in part of the news corpus also allows for comparison to an upper bound based on inter-annotator agreement.

In Chapter 4, relation identification is explicitly evaluated for the first time in the context of GRE, comparing window-based models defined in terms of intervening tokens to a novel model defined in terms of syntactic governor-dependency paths. Results suggest that a combined approach should be preferred as it is better in terms of recall and accuracy is shown to be comparable across domains. Importantly for applications of GRE, analysis demonstrates that at least 75% of false positive relations are actually implicit relationships that are not part of the gold standard relation type schemas.

In Chapter 5, relation characterisation experiments compare a number of similarity models, parametrised by feature set and dimensionality reduction technique. A novel feature set is introduced for the task based on syntactic features from governor-dependency parses. Comparison of dimensionality reduction techniques shows that a similarity model based on latent Dirichlet analysis (LDA) – a probabilistic generative approach – successfully incorporates a larger and more interdependent feature set than an unreduced model and a model based on singular value decomposition (SVD). LDA offers as much as a 34.5% reduction in the error rate when compared to SVD. And, while not always significant, it achieves higher f-scores than other approaches on five out of six evaluation settings. Taken together with the superior interpretability of the probabilistic generative approach, this motivates the use of LDA in the application here.

Finally, in Chapter 6, this thesis explicitly demonstrates the utility of relation discovery by incorporating GRE models as a conceptual representation for extractive text summarisation. This is evaluated with respect to a standard representation from the literature that uses weighted word tokens to represent sentence semantics. Results demonstrate that the GRE-based representation leads to improvements over the word token baseline. Analysis suggests that different representations do well on different types of summaries and that system combination will thus lead to improved performance.

Chapter 2

Literature Review

A vast amount of usable electronic data is in the form of unstructured text. The relation extraction task aims to identify useful information in text and recode it in a structured format that understood by machines. However, adapting conventional relation extraction systems to new domains or tasks requires significant effort from annotators and developers. This motivates an approach based on generic techniques that achieve comparable accuracy when transferred, without modification, across domains and tasks. A detailed comparison of previous generic approaches and evaluation frameworks highlights shortcomings with respect to task formalisation, modelling and extrinsic evaluation.

2.1 Introduction

Relation extraction (RE) can be addressed using supervised, bootstrapping or generic approaches. These have advantages and disadvantages which will be discussed in detail in the next section. One way to characterise them 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 characterise these approaches is in terms of the ontology creation problems they address, i.e. whether they address 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 tradeoff is in terms of accuracy, where generic approaches suffer when compared to supervised and bootstrapping approaches. However, as discussed in the applications and future work sections of this thesis (e.g., Chapters 6 and 7), generic approaches have high utility in terms of developing cheap components for applications, initialisation of semi-supervised bootstrapping and automated data exploration and visualisation.

This thesis develops generic approaches for RE. The task is referred to as generic relation extraction (GRE) because it is labour and cost effective and because it makes no assumptions about the data (i.e., the relation schema is learnt rather than being specified as part of the problem formulation). Previous approaches to GRE come from two distinct literatures, both of which are reviewed in this chapter. The first addresses the generic relation identification task (also known as relation mining), which aims to identify pairs of associated entities from text. And the second addresses the generic relation characterisation task (also known as relation discovery), which aims to characterise pairs of associated entities (i.e., annotate them with a label that describes the type of association). A detailed comparison of these approaches motivates the work in the rest of this thesis, which introduces 1) a rigorous intrinsic evaluation, 2) an assessment of utility with respect to a concrete application and 3) novel state-of-the-art models for GRE.

The remainder of this chapter contains a review of RE and adaptation. First, Section 2.2 contains a discussion of the origins of the RE task within the context of the history of information extraction. Next, Section 2.3 contains a discussion of different RE approaches which are characterised in terms of domain adaptation costs and whether they address ontology instantiation or learning. This is based on the basic division between supervised, bootstrapping, and generic approaches.

2.2 The Relation Extraction Task

2.2.1 The Information Extraction Framework

There are a number of aspects of the literature on natural language that address relations such as syntactic, semantic and pragmatic relations in theories of syntax and semantics (e.g., van Valin, 2006). Within the NLP literature, these correspond to tasks

such as phrase structure parsing (e.g., Jurafsky and Martin, 2000, chapters 9-12), dependency parsing (e.g., Nivre et al., 2007), semantic role labelling (e.g., Carreras and Màrquez, 2005), semantic interpretation (e.g., Bos, 2005), and discourse parsing (e.g., Marcu, 2006). There is also a literature on relations within philosophy, which focuses on the semiotic aspects of relationships (e.g., Peirce, 1870).

The current work is driven by the information extraction (IE) framework which has the practical goal of extracting structured information from natural language (e.g., Turmo et al., 2006). IE as a task was formalised largely in the context of the Message Understanding Conference (MUC) shared tasks (e.g., MUC-5, 1993; MUC-6, 1995; MUC-7, 1998) and more recently the ACE and BioCreAtIvE shared tasks (e.g., Doddington et al., 2004; Hirschman et al., 2004). IE is actually a collection of different sub-problems, whose core tasks include named entity recognition (NER), relation extraction (RE), and event extraction (EE). NER is the task of identifying and labelling named objects in text such as people, organisations, and locations. RE is the task of identifying associations between two entities, e.g. partner, subsidiary. EE is the task of identifying activities or occurrences such as mergers and acquisitions, airline crashes, and terrorist activities. The sub-problems of IE are generally considered to be incremental in nature, where event extraction builds on relation extraction and relation extraction builds on named entity recognition. Other IE tasks include coreference resolution and temporal analysis. In coreference resolution, entity mentions that refer to the same underlying entities are linked. Coreference resolution can be seen as a sub-task of NER or as a bridge between NER and subsequent IE tasks (Chinchor, 1998). In temporal analysis, event expressions are linked to related time expressions (e.g. Verhagen et al., 2007). This thesis focuses on RE.

2.2.2 Relation Identification and Characterisation

RE is often motivated by targeted intelligence needs such as keeping up-to-date information on companies as reported in the news (e.g., *PersonX* works for *OrganisationY*, *OrganisationY* owns *OrganisationZ*) or keeping up-to-date information on protein and gene interactions reported in the scientific literature (e.g., *GeneX* encodes *ProteinY*, *ProteinY* bind *ProteinZ*).[1] Here, the goal is to identify mentions of relations in text,

[1] A closely related task is labelling of semantic relations between nominals (e.g., Girju et al., 2007), which has the aim of identifying more general ontological relations between nominals such as those found in WordNet (Fellbaum, 1998) or Cyc (Matuszek et al., 2006). These relations include e.g. the cause-and-effect relation in the phrase "smile lines", the product-producer relation in the phrase "honey bee", and the content-container relation in "the apples in the basket". By contrast, relations in the

where a relation mention is defined as follows:[2]

> *A relation mention is a predicate ranging over two arguments, where an argument represents concepts, objects or people in the real world and the relation predicate describes the type of stative association or interaction that holds between the things represented by the arguments.*

In the news data used in the experimental chapters here, for example, there is a BUSINESS relation between "David Murray" and "Amidu Berry" in the text snippet "David Murray hired Amidu Berry." Saying that a relation is stative means that it describes a state of association or interaction that persists through time (though it may have a beginning and end point). This is in contrast to events which are generally more discrete in nature, describing things that happen or occur (e.g., Pustejovsky et al., 2003). In the example sentence above, for instance, a hiring event is described where David Murray is the one doing the hiring and Amidu Berry is the one being hired. The BUSINESS relation (that has a beginning point in time marked by the hiring act), by contrast, is an association between David Murray and Amidu Berry that persists through time (most likely until some other event like termination of the project or contract occurs). While relation may actually be argued to be a subclass of event, it is nevertheless advantageous to address relation extraction as an atomic task that is both tractable and useful.

While full definition of terminology is left for Chapter 3 (where the specific task addressed by this thesis is described), it is useful to include some basic terminology before going into the detailed review of GRE-related literature. First, instances of textual references to a relation are referred to here as *relation mentions*. Likewise, instances of textual references to an entity are referred to as *entity mentions*. Finally, labels used to describe the taxonomic class of relations and entities are referred to as *relation types* (e.g., "family" in the example above) and entity types (e.g., "person" in the example above).

Figure 2.1 contains a pipeline representation of the two main sub-tasks of RE: relation identification and relation characterisation. Input consists of natural language documents containing e.g. unstructured text or speech transcripts. These documents are first fed to the relation identification system, which identifies pairs of relation-forming entity mentions (e.g., "David Murray" and "Amidu Berry" in the example

information extraction task are generally defined as predicates over entity mentions in text that can be grounded to specific entities in the world.

[2]The specific notions of what constitutes a relation mention are derived from the data sets used for evaluation. These are described in Section 3.3. For more details, it is also useful to consult the annotation guidelines (LDC, 2004c, 2005b; Ginter et al., 2007).

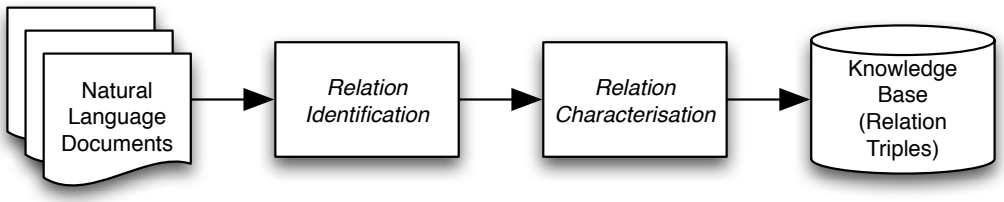


Figure 2.1: *Overview of main relation extraction sub-tasks.*

above). Next, the relation characterisation system annotates the entity mention pairs with a label describing the relation type (e.g., the BUSINESS label describing the relation between "David Murray" and "Amidu Berry" in the example above). Not all approaches to RE consist of separate modules for relation identification and characterisation, however it is a useful distinction for evaluation and for synthesising previous approaches to GRE. In the next Chapter, these sub-tasks are described in detail in the context of GRE.

2.3 Relation Extraction and Adaptation

2.3.1 Supervised Approaches

Conventional approaches to RE are generally based on either rule engineering or supervised machine learning, both of which are incur substantial development costs. Rule engineering requires extensive effort from a language engineer, who must be expert in the target domain and also be trained to develop an extraction grammar. For example, a rule that captures the relation between "David Murray" and "Amidu Berry" in the sentence "David Murray hired Amidu Berry" might say that two PERSON entity mentions with an intervening verb denoting a business transaction (e.g., "hire") constitute a BUSINESS relation. In supervised machine learning, an annotated corpus is used to train a new system. For example, the news corpus used here contains the example sentence above where the entity and relation mentions have been marked by a human annotator. A supervised machine learning system would extract features of the entity mentions (e.g., word tokens, type) and the surrounding context (e.g., word tokens, grammatical information) and learn a mapping from features to relation types. While approaches to RE based on supervised machine learning were motivated by the

expense of adapting rule-based systems to new domains or tasks, they require substantial effort from annotators, who must be expert in the target domain. In addition, the reality is that they still generally require effort from a language engineer for feature engineering and tuning of algorithm parameters.

In terms of adaptation, rule-based systems always require re-writing the extraction grammar, though the actual amount of effort varies depending on how much a new relation schema differs from that of an existing system. For supervised machine learning, domain adaptation may be achieved by transfer learning, where existing annotation is used to inform an extraction task in a new domain. Chu et al. (2002) and Daumé III (2007) describe approaches to transfer learning where models are adapted to the new domain using a small amount of annotated data. And, Blitzer et al. (2006) describe an approach that identifies correspondences between features in the source and target domains (Blitzer et al., 2006). Another approach to adaptation that does not require similar schemas is active learning, where a supervised classifier is trained on a small seed set of labelled data for a new domain after which classification uncertainty is exploited to select examples for annotation that are most useful for the learning algorithm (e.g., Seung et al., 1992; Cohn et al., 1996; Hachey et al., 2005; Zelenko et al., 2005). However, supervised adaptation approaches always require some amount of rule engineering or annotation. Relation extraction suffers particularly in this respect as it is a complex task, which means e.g. less annotation output per hour of labour. Furthermore, supervised adaptation approaches require that the relation schema be known in advance. Therefore, they are not appropriate e.g. for exploratory relation extraction on new domains or sub-domains where the schema is not known.

2.3.2 Bootstrapping Approaches

The expense of conventional supervised approaches for natural language processing has also motivated recent work on partially supervised methods for bootstrapping RE in new domains. These typically exploit large amounts of unlabelled data to bootstrap a wide-coverage system and can be divided into three main types of approaches, which are characterised below by what type of gold standard data they require for initialisation.

The first type of approach requires labelled training data for a new domain like the supervised machine learning approaches, but uses a relatively small amount. A wide-coverage system is then bootstrapped through an iterative process of learning and

automatic annotation of new training data. Co-training is an instance of this type of approach where two relation extraction systems are trained based on distinct views of each example (Blum and Mitchell, 1998). Zhang (2004) present co-training approach to the relation characterisation sub-task of RE, where multiple views are automatically created by random projection from the original feature space. Hassan et al. (2006) present another instance of this type of approach sometimes referred to as self-training, where a trained classifier is used to annotate unlabelled data followed by a ranking process that identifies data points to add to the training data.

The second type of approach requires a small amount of example entity pairs for a specific type of relation. A wide-coverage system is then bootstrapped through an iterative process of 1) identifying texts where the example entities occur together and using these texts to induce extraction patterns and 2) using the induced extraction patterns to identify new entity pairs between which the target relation holds (e.g., Brin, 1998; Riloff and Jones, 1999; Agichtein and Gravano, 2000; Agichtein, 2006; Tomita et al., 2006). For example, Brin (1998) bootstrap an extraction system for identifying AUTHOROF relations between book titles and people starting from a seed set of five book-people pairs such as "The Robots of Dawn" and "Isaac Asimov".

The third type of approach requires only a group of documents classified as relevant or non-relevant to a particular extraction task, though these can be derived from an information retrieval system given a query (e.g., Sudo et al., 2003). The distinction between relevant and non-relevant is used to assign a weight that balances pattern frequency in relevant documents against frequency across relevant and non-relevant documents. Greenwood and Stevenson (2007) apply this approach to ranking extraction patterns for a relation identification task, where the goal of the system is to identify pairs of entity mentions that are part of the same event.

These bootstrapping approaches are designed to learn with minimal engineering, requiring only small sets of seed data and are appropriate for situations in which resources are minimal. However, like the supervised approaches, they require that the relation schema be known in advance and are therefore not appropriate e.g. for exploratory relation extraction on new domains where the schema is not known.

2.3.3 Generic Approaches

Another way to address adaptation is by moving from fully supervised approaches (e.g., machine learning techniques that require annotated data for each new domain)

to generic approaches that do not require any annotation or parameter tuning when moving to new domains. This addresses the shortcomings discussed in the previous two sections with respect to both cost and assumptions about the data. As regards cost, generic approaches can be developed with reference to one domain and achieve comparable accuracy when transferred, without modification of model parameters, to other domains. This will be demonstrated in this thesis for the GRE task. As regards assumptions, generic approaches do not require the relation schema to be specified as part of the problem formulation and can thus be used for exploratory relation extraction in new domains where the schema is not known. In other words, they can be used for learning the structure of ontologies as well as learning how to instantiate them. Generic approaches have been explored recently for a number of NLP tasks including coreference resolution (e.g., Haghighi and Klein, 2007), part-of-speech tagging (e.g., Johnson, 2007; Goldwater and Griffiths, 2007) and parsing (e.g., Smith, 2006; Bod et al., 2003; Klein, 2005).

Generic approaches have also been applied to RE. Conrad and Utt (1994) describe an approach to identifying associated pairs of named entities from a large corpus using statistical measures of co-occurrence. This task will be referred to as generic relation identification in this thesis, but is also known as relation mining in the literature. In more recent work, Hasegawa et al. (2004) describe an approach to characterising co-occurring named entity pairs by relation type. This task will be referred to as generic relation characterisation in this thesis, but is also known as relation discovery in the literature. This uses automatic clustering to induce a partition over the relation-forming entity pairs and cluster labelling techniques to annotate clusters with a relation type. In other related work, Filatova and Hatzivassiloglou (2003) and Filatova et al. (2006) describe generic approach that is oriented more towards event extraction. The following section contains a detailed review of the GRE literature, focusing on previous approaches to modelling and evaluation.

2.4 Review of the GRE Literature

This section contains a detailed survey of related work in terms of modelling and evaluation for the two main sub-tasks of GRE: relation identification and characterisation. For both tasks, models from previous approaches are characterised in terms of a set of common parameters. This characterisation is summarised in overview tables that allow easy comparison of the different approaches. Details of each approach are pre-

sented in dedicated sub-sections in chronological order of publication. The evaluation performed for previous approaches is also characterised in terms of a set of common parameters. Again, this is summarised in overview tables to allow easy comparison while reading through the detailed sections addressing individual approaches.

Previous work has focused on intrinsic evaluation, where GRE is evaluated as a task on its own. Most of these evaluations use a common definition of accuracy such as precision (percent of system answers that are correct), recall (percent of possible correct answers among the system results) or f-score (harmonic mean of precision and recall). For formal definitions and further discussion of intrinsic evaluation measures, refer to Section 3.4. Another way to evaluate is by embedding a system in another application or task and quantifying its impact on the the application or task. This is referred to as extrinsic evaluation (e.g., Sparck Jones and Galliers, 1996). Extrinsic evaluation will be discussed further at the end of this chapter and is the primary focus of Chapter 6 of this thesis.

2.4.1 Generic Relation Identification

The seminal work on discovery of novel entity associations is due to Swanson (1986), who introduced a system that allows a user to state a hypothesis about two items A and C being related. The system searches MEDLINE for papers describing A and for papers describing C, using these two construct a list of words and phrases common to both sets of papers. The system was used to propose fish oil as a novel treatment for Raynaud's disease (a circulatory disorder restricting blood-flow to the extremities), a hypothesis which was later confirmed through wet lab experiments. Recent work, by contrast, has focused on more general solutions that do not necessarily require a hypothesis or query as input. Here, the generic relation identification (GRI) task aims to identify pairs of associated entities from text.

Table 2.1 contains an overview of modelling approaches from the GRI literature. The first column (Citation) contains the reference to the authors of the approach. The first four rows in the table correspond to approaches that focus on the GRI task while the last four rows correspond to approaches that focus on the GRC task which is addressed below in Section 2.4.2. The second column (Co-occur Window) describes the window for identifying entity mention pairs. The third column (Constraints) describes any additional constraints placed on entity mention pairs. And the fourth column (Weighting) contains the weighting scheme(s) used for ranking entity pairs. Where the

Citation	Co-occur Window	Constraints	Weighting
Conrad and Utt (1994)	W/in 25, 100 words	NA	PMI, ϕ^2
Jenssen et al. (2001)	Document	NA	C
Smith (2002)	Sentence	NA	PMI, ϕ^2, χ^2, G^2, C
Filatova and Hatz. (2003)	Sentence	Verbal connector	Pr
Hasegawa et al. (2004)	W/in 5 words	NA	NA
Chen et al. (2005)	Gold, W/in 10	NA	NA
Zhang et al. (2005)	Same sentence	Spanning parse	NA
Chen et al. (2006)	Gold	NA	NA

Table 2.1: *Overview of modelling approaches from the generic relation identification literature. Columns correspond to the author (Citation), the window for identifying entity mention pairs (Co-occur Window), constraints on entity mention pairs (Constraints) and weighting schemes for entity mention pairs (Weighting).*

word "Gold" is present in the Co-occur Window column, the authors use gold standard entity mention pairs from annotated RE data. Weighting schemes include frequency (C), probability (Pr), chi-squared (χ^2), phi-squared (ϕ^2), pointwise mutual information (PMI) and log-likelihood ratio (G^2). The discussion in the rest of this section will provide further details.

Table 2.2 contains an overview of evaluation frameworks from the GRI literature. The first column (Citation) contains the reference to the authors of the approach. The GRC systems (Hasegawa et al., 2004; Chen et al., 2005; Zhang et al., 2005; Chen et al., 2006) are not included here because they only perform evaluations of relation characterisation performance, which are discussed below in Section 2.4.2. The second column (Data) contains the name of the data set used. The third (NER) and fourth (Coref) columns describe the named entity recognition that was used and the approach to coreference. In the Data, NER and Coref columns, "Internal" indicates that the authors use internal, non-public resources. In the NER and Coref columns, "Index" indicates that NER and coreference are performed using a term matching procedure (described in Section 2.4.1.2 below) and "Overlap" indicates that coreference is based on the overlap between verbal connector words of two entity pairs (described in Section 2.4.1.4 below). The fifth column (Gold) describes what was used as a gold standard reference for evaluation. The sixth column (Sub-Doms) lists the entity pair sub-domains used for evaluation. These include pairs consisting of two COMPANY entities (C-C), pairs

Citation	Data	NER	Coref	Gold	Sub-Doms	Eval
Conrad and Utt (1994)	WSJ	Internal	None	Manual	C-C, C-P, P-P	*P*, *R*
Jenssen et al. (2001)	PubMed	Index	Index	Curation	G-G	*R*
Smith (2002)	Internal	Internal	Internal	Curation	L-D	*MRR*
Filatova and Hatz. (2003)	TDT2	IdentiFind	Overlap	Manual	Combined	*P*

Table 2.2: *Overview of evaluation frameworks from the generic relation identification literature. Columns correspond to the author (Citation), the data set (Data), the named entity recognition used (NER), the approach to coreference resolution (Coref), the type of gold standard reference information (Gold), the entity pair sub-domains evaluated (Sub-Doms) and the evaluation measure (Eval).*

consisting of one COMPANY entity and one PERSON entity (C-P), pairs consisting of two PERSON entities (P-P), pairs consisting of two GENE entities (G-G) and pairs consisting of one LOCATION entity and one DATE entity (L-D). "Combined" indicates that the evaluation does not consider entity pair sub-domains individually. Finally, the seventh column (Eval) contains the evaluation measure, which can include precision (*P*), recall (*R*) and or mean reciprocal rank (*MRR*). The discussion in the rest of this section will provide further details.

2.4.1.1 GRI: Conrad and Utt (1994)

Conrad and Utt (1994) present seminal work on mining pairs of entities from large text collections. The system uses statistical measures of association to rank named entity pairs by presumed importance based on co-occurrence. Conrad and Utt 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. They do specify a minimum co-occurrence threshold of 2. Conrad and Utt use two statistical association measures for ranking entity pairs: pointwise mutual information (*PMI*) and phi-squared (ϕ^2). *PMI* compares the probability of observing two words together with the probability of observing them independently. ϕ^2 is a statistical test for analysing deviance from expectation in enumeration data. The authors choose ϕ^2 for their experiments as it tends to favour high-frequency associations.

Conrad and Utt (1994) perform a manual evaluation using Wall Street Journal (WSJ) data (Harman, 1992), using articles from 1987 and 1991 respectively for development and testing. NER is performed automatically using an internal tool. No coreference resolution is performed. Accuracy is calculated using precision and recall (see Section 3.4.1). The authors use three entity pair sub-domains for evaluation: 1) pairs consisting of a COMPANY entity and a PERSON entity (C-P), 2) pairs consisting of two COMPANY entities (C-C) and 3) pairs consisting of two PERSON entities (P-P). To calculate precision, fifteen COMPANY and fifteen PERSON entities were chosen. Next, the system was used to generate entity associations from the test data for the chosen entities and the associations were then manually classified as true or not true. To calculate recall, two PERSON entities and two COMPANY entities were chosen. Next, a retrieval engine was used to collect all documents in the collection containing the chosen entities and the documents were then manually annotated for entity associations. Combining precision and recall gives balanced f-scores of 79.4, 77.7 and 86.3 respectively for the C-C, C-P and P-P entity pair sub-domains. This is not compared to any lower or upper performance bounds. Also, no criteria are given for the choice of entities for evaluation so it is not clear whether these are random (i.e. representative of the distribution in the input data) or not (i.e. skewed towards high or low frequency entities). Furthermore, because the evaluation is performed manually and is based on the output of their system, it cannot be re-created for the sake of comparison.

2.4.1.2 GRI: Jenssen et al. (2001)

Jenssen et al. (2001) describe a similar methodology that uses co-occurrence in the biomedical literature to create a weighted network of gene relations. This builds on previous biomedical text mining work aimed at building systems that can automatically discover relationships that can be formulated as meaningful research hypotheses, which can subsequently be tested in biological wet lab experiments (e.g., Swanson, 1986; Smalheiser and Swanson, 1998; Blaschke et al., 1999; Stapley and Benoit, 2000; Rindflesch et al., 2000). Jenssen et al. use documents (i.e., titles and abstracts) to calculate raw co-occurrence counts, which are used to weight gene entity pairs. They do not consider the use of statistical measures of association to account for chance co-occurrence.

Jenssen et al. (2001) perform an evaluation using PubMed[3] data. NER is performed

[3] http://www.ncbi.nlm.nih.gov/PubMed/

by a simple indexing procedure that searches texts for gene names contained in a large gazetteer, which is also used to perform coreference resolution by mapping gene mentions to the primary symbol associated with the underlying real-world entity. Accuracy is calculated by computing recall (see Section 3.4.2) with respect to manually curated entity pairs from two databases, where system pairs are considered true if they are present in the databases. The authors consider one entity pair sub-domain for evaluation: pairs consisting of two GENE entities (G-G). The system achieves recall scores of 0.51 and 0.45 respectively for the DIP and OMIM databases. With respect to perfect performance (recall of 1.00), this represents reductions in the error rates over the baseline (random generation of interacting protein pairs) of 46.8% and 44.5%. While Jenssen et al. (2001) formulate a sound experimental procedure using publicly available resources, they do not consider precision, meaning that a trivially optimal solution could be achieved by a system that proposes all possible pairs of entities. Also, they do not compare to an upper bound on accuracy.

2.4.1.3 GRI: Smith (2002)

Other statistical measures of association can also be used. Smith (2002) looks at chi-squared (χ^2) and log-likelihood ratio (G^2) in addition to the *PMI* and ϕ^2. χ^2 is the unnormalised version of ϕ^2 and is included for completeness. The introduction of G^2 is motivated by Dunning (1993), who argues that measures like *PMI* and z score are unreliable where counts are low. Smith (2002) performs an evaluation using an internal corpus of nineteenth century historical documents focusing on British and American history. NER and coreference are performed using internal tools. Accuracy is computed with respect to a curated resource, which contains expert assessments of the severity of battles in the American civil war (Dyer, 1960). The accuracy measure used is the mean reciprocal rank (*MRR*), i.e. the inverse of the rank of the first correct answer. The authors consider one entity pair sub-domain for evaluation: pairs consisting of a LOCATION entity and a DATE entity (L-D). G^2 was compared to the next best system which simply uses raw frequency counts to rank entity pairs. Results show no statistically significant difference for pairs with frequency greater than or equal to five. However, G^2 was found to perform significantly better when low-frequency pairs were included. Smith (2002) does not compare to lower or upper bounds. And, because his methodology relies on an internal tools and an internal corpus, the evaluation framework cannot be re-created for the sake of comparison.

2.4.1.4 GRI: Filatova and Hatzivassiloglou (2003)

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 atomic parts of event descriptions and they add a constraint requiring that a verbal 'connector' (i.e., a verb or a noun that is a WordNet hyponym of *event* or *activity*) be present in the intervening token context between the entity mentions. The authors present a limited evaluation based on manual analysis of the system output that uses the IdentiFinder system for NER.[4] They use string matching for automatic coreference and also adopt a secondary approach to that considers two entities B and C to be equivalent for the purpose of their relationship to a third entity A if the connectors that occur between A and B have at least 75% overlap with the connectors that occur between A and C. Results suggest that the system achieves reasonable precision. The evaluation does not, however, address recall and it does not compare the system to any lower or upper bounds on accuracy. Follow up work (see Chapter 6) describes a more rigorous extrinsic evaluation based on extractive text summarisation.

2.4.2 Generic Relation Characterisation

The generic relation characterisation (GRC) task aims to characterise pairs of associated objects, annotating them with a relation type extracted from the textual context. The GRC literature approaches the relation characterisation task as a clustering problem, where the goal is to induce a partition over entity pairs that groups them by relation type. Then, cluster labelling techniques are applied to annotate clusters with a relation type label.

Table 2.3 contains an overview of modelling approaches from the GRC literature. The first column (Citation) contains the reference to the authors of the approach. The second column (Features) describes the features used to represent entity pair instances. Here, CC refers to constituent chains derived from a phrase structure parser (see Section 2.4.2.4). The third column (Similarity) contains the similarity measure used to compare entity pair instances. The fourth column (Mod Order Sel) refers to the approach to model order selection, i.e. identifying the number of clusters. And the fifth column (Clustering) describes the clustering algorithm used. The discussion in the rest of this section will provide further details.

[4] http://www.bbn.com/technology/data_indexing_and_mining/identifinder

Citation	Features	Similarity	Mod Order Sel	Clustering
Hasegawa et al. (2004)	Intervening words	Cosine	Hier clust w/ sim thresh	Agglomerative (complete-link)
Chen et al. (2005)	Intervening words	Cosine	Stability-based resampling	k-means
Zhang et al. (2005)	Smallest parse fragment spanning ents	Tree kernel	Hier clust w/ sim thresh	Agglomerative (group ave)
Chen et al. (2006)	Words; POS, Ent & Chunk types; CC	Cosine	Spectral order detection	Spectral

Table 2.3: *Overview of modelling approaches from the GRC literature. Columns correspond to the author (Citation), the feature set (Features), the similarity measure (Similarity), the approach to model order selection (Mod Order Sel) and clustering approach (Clustering).*

Table 2.4 contains an overview of evaluation frameworks from the GRC literature. The first column (Citation) contains the reference to the authors of the approach. The second column (Data) contains the name of the data set used. The third (NER) and fourth (Coref) columns describes the named entity recognition that was used and the approach to coreference. "Gold" indicates that the authors used gold standard relation-forming entity mention pairs as input to the GRC evaluation. "Str Eq" indicates that coreference resolution is based on string equality. The fifth column (Instance) describes the instance level for clustering. "Types" indicates that clustering instances include the concatenated contexts of all entity mention pairs for two given entities and "Tokens" indicates that every individual mention of an entity pair is considered a separate clustering instance. "UNK" indicates that the authors do not specify the instance level. The sixth column (Gold) describes what was used as a gold standard reference for evaluation. "Manual" indicates a manual evaluation of the system output and "Annotation" indicates an evaluation with respect to an annotated gold standard corpus. The seventh column (Sub-Doms) lists the entity pair sub-domains used for evaluation. These include pairs consisting of a PERSON entity and a GEO-POLITICAL entity (P-G), pairs consisting of two ORGANISATION entities (O-O), pairs consisting of a PERSON entity and an ORGANISATION entity (P-O) and pairs consisting of an ORGANISATION entity and a GEO-POLTICAL entity (O-G). "Combined" indicates that the evaluation does not consider entity pair sub-domains individually. Finally, the eighth column (Eval) contains the evaluation measure. This is either the many-to-one f-score ($F_{n:1}$)

Citation	Data	NER	Coref	Instance	Gold	Sub-Doms	Eval
Hasegawa et al. (2004)	NYT	OAK	Str Eq	Types	Manual	P-G, O-O	$F_{n:1}$
Chen et al. (2005)	ACE	Gold	None	UNK	Annotation	P-O, O-G, O-O	$F_{1:1}$
Zhang et al. (2005)	NYT	OAK	None	Tokens	Manual	P-G, O-O	$F_{n:1}$
Chen et al. (2006)	ACE	Gold	None	UNK	Annotation	Combined	$F_{1:1}$

Table 2.4: *Overview of evaluation frameworks from the generic relation characterisation literature. Columns correspond to the author (Citation), the data set (Data), the named entity recognition used (NER), the approach to coreference resolution (Coref), the instance level for clustering (Instance), the type of gold standard reference information (Gold), the entity pair subdomains evaluated (Sub-Doms) and the evaluation measure (Eval).*

or the one-to-one f-score ($F_{1:1}$), which are defined in Chapter 3. The discussion in the rest of this section will provide further details.

2.4.2.1 GRC: Hasegawa et al. (2004)

Hasegawa et al. (2004) introduce the task of GRC (which they refer to as relation discovery) and describe it in terms of the high-level algorithm in Figure 2.2. The first and second steps perform pre-processing. The first step is NER, where entity mentions are identified. Hasegawa et al. use an off-the-shelf tagger called OAK with an extended hierarchy of 150 entity types (Sekine, 2001). The second step corresponds directly to the GRI task from the previous section (2.4.1), where relation-forming entity pairs are extracted. Hasegawa et al. (2004) use a simple approach 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. Furthermore, they do not say e.g. whether stop words in the intervening context are considered or whether other entity mentions are allowed. Hasegawa et al. also do not explicitly evaluate the accuracy of their approach to relation identification. The third through fifth steps constitute the core GRC task. Hasegawa et al. (2004) set the trend for the literature in focusing on Steps 3 and 4, which constitute the fundamental modelling problems of any clustering task. The following description of their framework serves to introduce the GRC task.

The third step is concerned with building a similarity matrix to be input to the clustering algorithm. Hasegawa et al. (2004) save the intervening tokens from each pair of entity mentions to be used as features for the clustering algorithm. Pre-processing

1	Identify named entity mentions
2	Extract pairs of co-occurring named entities
3	Build similarity matrix
4	Cluster named entity pairs
5	Label clusters

Figure 2.2: *Overview of Hasegawa et al. (2004) approach to relation characterisation.*

includes the removal of stop words from intervening contexts.[5] Hasegawa et al. also ignore what they term parallel expressions in intervening context, which consist of tokens used to punctuate lists (i.e., ",.*,", "and" and "or"). They also strip datelines (e.g., "WASHINGTON (AP) –" at the beginning of a document) from their corpus to avoid erroneous relations involving entity mentions found here. Next, they match entity mentions based on string equality (i.e., entity mentions with the same surface string are considered to refer to the same underlying entity) and combine intervening token contexts of all matching pairs. For example, "*OrganisationA* offered to buy *OrganisationB*" and "*OrganisationA*'s proposed purchase of *OrganisationB*" would produce a combined context for the $<OrganisationA, OrganisationB>$ tuple consisting of the following stemmed versions of the intervening tokens: "offer", "to", "buy", "'s", "propose" and "purchase". The authors then derive weights for word tokens and compute similarity between the resulting feature vectors for entity pair contexts. Hasegawa et al. use a *tf*idf* weighting scheme. This aims to balance a term's frequency (*tf*) in a given context with how common it is across contexts (*idf*) (Spärck Jones, 1972; Salton and McGill, 1986). Sparse vectors are removed based on a minimum threshold for vector norm values, which the authors set to 10. Then, similarity between weighted feature factors is computed using cosine (defined in Section 5.2.1.2).

In the fourth step, the actual clustering is performed, resulting in a partition that is intended to group entity pairs by their relation type. Hasegawa et al. (2004) adopt hierarchical clustering as it is not known in advance how many clusters there should be and they adopt the complete-link criterion function because it is conservative in making clusters. Complete-link (e.g., Tan et al., 2005, p517) measures the similarity of two clusters by the minimum similarity between feature vectors from each cluster.

[5] Hasegawa et al. (2004) define stop words as tokens with corpus frequency less than 3 or greater than 100000.

This is used in agglomerative (i.e., bottom-up) hierarchical clustering to determine which two clusters are the most similar and thus should be merged next. Hasegawa et al. (2004) control the number of clusters in the final solution by setting a minimum threshold on the similarity required to merge two clusters. For their final evaluation, the authors set the threshold to what they describe as a value just above zero.

Hasegawa et al. (2004) perform an evaluation of the first four steps of their system using one year of newswire data from the New York Times (NYT). Accuracy is computed based on a manual human classification of the entity pairs extracted by the system in the second step, which is subsequently used to compute a balanced f-score ($F_{n:1}$) based on a many-to-one mapping between clusters and gold standard classes (see Section 3.4.3.3). They chose two entity pair sub-domains for evaluation: 1) pairs consisting of a PERSON entity and a GEO-POLITICAL entity (P-G)) and 2) pairs consisting of two COMPANY entities (C-C). For both of these sub-domains, the authors take the output of the automatically extracted pairs of co-occurring entities and classify them manually, serving to create a gold standard partition over the data. Hasegawa et al. report f-scores of 80 and 75 respectively for the P-G and C-C entity pair sub-domains. They do not compare this to any performance bounds. Furthermore, because their methodology relies on the output of their relation identification, the evaluation cannot be re-created for the sake of comparison.

Finally, in the fifth step, labels are chosen for the clusters to serve as relation type annotation. Hasegawa et al. (2004) simply select the context word tokens with the highest frequency. Specifically, they weight the descriptiveness of a word token w_i by calculating the number of times two entity pairs in the cluster both have w_i in their context. This is normalised by the total number of pairwise contacts between entity pairs in the cluster to give a value between 0 and 1. The authors do not explicitly evaluate the automatically derived labels.

2.4.2.2 GRC: Chen et al. (2005)

Chen et al. (2005) identify two limitations of the Hasegawa et al. (2004) approach to the GRC task. First, they note that the similarity threshold method for identifying the number of clusters is not a very good solution as it is not guaranteed to generalise. Second, the authors suggest that it is not sufficient for cluster labels to be descriptive; they should be discriminative as well. That is, cluster labels should differentiate between clusters as well as being indicative of cluster content. This is described in detail below.

The first contribution of the Chen et al. (2005) approach is to demonstrate the use of partitional (k-means) clustering with automatic model order selection (i.e., automatically determining the number of clusters). They employ a criterion function using resampling-based stability analysis (Lange et al., 2003), which had been used previously for document clustering (Niu et al., 2004). The authors describe an approach that evaluates all possible numbers of clusters k over a prespecified range and chooses the k that maximises the criterion function. The criterion function is computed by 1) randomly permuting the clustering instances, 2) clustering a subset (90% of full data set) of the permuted instances, and 3) measuring consistency of the resulting clustering with result to the clustering over the full data set. Chen et al. (2005) measure consistency by computing the purity (see Section 3.4.3.3) of the clustering solution over the resampled data with respect to the clustering solution over the original data set and repeat the process an unspecified number of times.

Chen et al. also move towards formalisation of the task by introducing the use of a gold standard information extraction corpus for evaluation. They adopt the data from the Automated Content Extraction (ACE) shared tasks sponsored by the US National Institute of Standards and Technology.[6] This allows them to isolate the performance of the clustering component by using gold standard relations, filtered by the number of intervening word tokens. This also suggests the possibility of comparative evaluation with other approaches given a more detailed specification of the edition of the data used and any pre-processing. Accuracy is computed with respect to the partition defined by the gold standard relation type annotation. The authors use a balanced f-score ($F_{1:1}$) based on a one-to-one mapping between clusters and gold standard classes (see Section 3.4.3.4).

Chen et al. (2005) evaluate on three entity pair sub-domains: 1) pairs consisting of a PERSON entity and an ORGANISATION entity (P-O), 2) pairs consisting of an ORGANISATION entity and a GEO-POLITICAL entity (O-G), and 3) pairs consisting of two ORGANISATION entities (O-O). For gold standard entity mention pairs occurring within ten word tokens of each other, the authors report f-scores of 39.3, 50.9 and 37.2 respectively for the P-O, O-G and O-O sub-domains. Taking the mean (macro average) across entity pair sub-domains, the approach achieves a 7.3 point increase over Chen et al.'s reimplementation of the Hasegawa et al. (2004) approach. With respect to perfect performance (f-score of 1.0), this represents a reduction in the error rate of 11.3%. Chen et al. do not evaluate on held out data and do not report whether their

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

results are statistically significant. Also, they do not compare to any upper bound.

The second contribution of the Chen et al. (2005) approach is to suggest that labels chosen to annotate clusters in the GRC task should differentiate between clusters as well as being descriptive. To achieve this, they propose the use of discriminative category matching techniques from the document classification literature (Fung et al., 2002). The motivation is the same as that for the *tf*idf* term weighting scheme and it is calculated in a similar way by combining a measure of how common a term is across clusters with the simple within-cluster term frequency used by Hasegawa et al. (2004). The authors select the two highest ranked labels to describe a cluster.

The authors also propose a method for automatic evaluation of the labelling task against gold standard relation labels. This relies in information content calculated using distributional information from a large corpus (Resnik, 1995) and the WordNet lexical ontology (Fellbaum, 1998). Information content (*IC*) of a term t_i is calculated as $-\log p(t_i)$. Thus, *IC* is high when the probability of encountering t_i is low, from which it follows that lower nodes in a concept taxonomy should have *IC* greater than higher nodes. Chen et al. apply an *IC*-based measure derived by Lin (1997), which is defined as:

$$Relatedness(l_i, l_j) = \frac{2 \times IC\left(lcs\left(l_i, l_j\right)\right)}{IC\left(l_i\right) + IC\left(l_j\right)} \tag{2.1}$$

where l_i and l_j are the labels that are being compared and $lcs(l_i, l_j)$ is the lowest common hypernym of l_i and l_j in a concept taxonomy (i.e., WordNet). The mean *Relatedness* scores for Chen et al. work out to 0.520, 0.571 and 0.581 respectively for the P-O, O-G and O-O sub-domains. While the authors do not report statistical significance tests, the discriminative labelling approach achieves a substantial improvement over a reimplementation of the Hasegawa et al. (2004) approach. The discriminative system achieves a mean increase of 0.215 points, or a 32.7% reduction in error rate with respect to perfect performance (*Relatedness* score of 1.0). However, relying on lexical resources like WordNet means that the technique does not port to domains where similar resources are not available.

2.4.2.3 GRC: Zhang et al. (2005)

Zhang et al. (2005) discuss one main contention with the earlier work of Hasegawa et al., namely that flat feature vectors based on intervening words are not sufficient for the GRC task. The authors propose a representation based on the smallest parse tree fragments spanning both entity mentions in a pair (spanning parse). The resulting

similarity function is based on the output of an automatic parser (Collins, 1999) and incorporates structural information about word tokens, parts-of-speech, phrase types, phrase heads, entity types and directionality in phrase structure trees. Like Hasegawa et al., the authors use hierarchical agglomerative clustering[7] and the number of clusters in the solution is controlled by a minimum threshold on the similarity required to merge two clusters. The Zhang et al. approach to relation identification also differs from Hasegawa et al. in that all pairs of entities in the same sentence are considered to be co-occurring given that there is a spanning parse, as opposed to only those that occur within five word tokens of each other.

Zhang et al. use the same evaluation framework as Hasegawa et al. (2004) apart from one significant change with respect to the instance level for clustering and evaluation. They argue against the simplifying assumption that pairs of relation mentions with coreferring entities always have the same relation and thus can be combined to create a single feature vector. Instead, they create a clustering instance from every mention of a entity pair. This can also be described in the language of semiotics (Peirce, 1933, Paragraph 537) as using entity pair tokens instead of entity pair types.

The authors evaluate on the same entity pair sub-domains as Hasegawa et al.: 1) pairs consisting of a PERSON entity mention and GEO-POLITICAL entity mention (P-G) and 2) pairs consisting of two COMPANY entity mentions (C-C). On high frequency entity pairs (co-occurring 30 or more times), they report many-to-one f-scores ($F_{n:1}$) of 87 and 80 respectively for P-G and C-C. Taking the mean across the entity pair sub-domains, the approach achieves a 4 point increase over a reimplementation of the Hasegawa et al. approach. With respect to perfect performance (f-score of 1.0), this represents a reduction in the error rate of 19.5%. Performance is substantially lower on less frequent pairs which is likely due in part to the fact that the tree-based similarity model is highly specified and does not include a simpler back-off representation to cope with sparsity or noise error propagation from the parser. Zhang et al. do not evaluate on held-out data and do not report whether their results are statistically significant.

The authors also propose a new approach to labelling clusters based on parse information. They argue that head words from the root nodes of the minimum spanning parse fragments are the best source for cluster labels. This is based on the notion of headedness in syntactic grammars, where a head is the word or category that gets prop-

[7]Zhang et al. (2005) use the group average criterion function instead of complete-link, based on performance. Group average measures the similarity between clusters as the average similarity between their members.

agated up a phrase structure tree. Another way of describing this notion is that a head is the main word associated with the root of a phrase or sentence and as such is the word that is described or specified by the non-head branches in a parse tree. Cluster labelling is not evaluated or compared to related work.

2.4.2.4 GRC: Chen et al. (2006)

Chen et al. (2006) propose another approach to the GRC clustering task. They incorporate a richer feature set than they used in their previous work on the task (Chen et al., 2005), which includes word tokens from the intervening context, from the context just before the first entity mention and just after the second entity mention, and from the entity mentions themselves. The feature set also includes type information for entity mentions and part-of-speech (POS) tags corresponding to all entity mention and context word tokens. Finally, they incorporate a number of grammatical features based on parse trees from the Charniak Parser (Charniak, 1999), including chunk phrase types and grammatical function information for all entity mention and context word tokens, and constituent chains. Constituent chains capture phrase embedding information information on the path from the root node of a phrase structure parse tree to the target leaf node (i.e., an entity mention).[8]

The authors suggest that previous approaches suffer from an inability to identify complex structures in the feature space. They apply spectral clustering to the problem, which performs an elongated k-means clustering[9] on the q eigenvectors with the highest eigenvalues computed from the Laplacian of the similarity matrix (e.g., Ng et al., 2002; Sanguinetti et al., 2005; von Luxburg, 2006). This spectral decomposition is essentially a dimensionality reduction technique for similarity matrices where the eigen decomposition provides a way to create a reduced representation based on the eigenvectors that explain the largest amount of similarity (Sanguinetti et al., 2005) and the Laplacian takes into account the different variance within the various clusters.[10]

The spectral clustering paradigm of Sanguinetti et al. (2005) also provides a simple incremental approach to model order selection (spectral order detection). This starts with $q = 2$ eigenvectors and continues performing k-means clustering initialis-

[8] See e.g. Zhang (2005) for details.

[9] Elongated k-means clustering down-weights distances along radial directions and penalises distances along transversal directions to account for the elongated nature of the clusters resulting from orthogonality of the reduced eigenvector space (Sanguinetti et al., 2005).

[10] In parallel work that forms part of Chapter 5 of this thesis, I also introduced dimensionality reduction to the GRC task (Hachey, 2006).

ing q clusters on the centres on the q eigenvectors and initialising an additional cluster on the origin. It iterates, incrementing q, until the additional cluster after performing elongated k-means is empty. Results suggest that this performs well on the GRC clustering task. The algorithm creates 21 clusters with the best feature set, where the gold standard number is 24.

The authors assess the clustering accuracy using the one-to-one f-score ($F_{1:1}$) from Chen et al. (2005) (see Section 3.4.3.4). However, they evaluate a single clustering task over all of the gold standard relations in ACE, i.e. they do not decompose the data into sub-domains based on entity types like previous approaches but rather evaluate on the full data with all entity pair sub-domains combined. The approach has one free feature weight scaling parameter that is tuned on a development test set. The authors report an f-score of 46.3, achieving a 12.6 point increase over a reimplementation of the Hasegawa et al. (2004). This represents a reduction in the error rate of 19.0% with respect to perfect performance and a reduction of 59.7% with respect to a supervised classifier using the same feature set. The authors do not report whether their results are statistically significant.

Chen et al. (2006) apply a very interesting technique to the clustering task, however their definition of the task does not take full account of the original motivation for the GRC task. Specifically, using entity mention word tokens and entity types as features may improve scores on the isolated clustering task, but placing two relations in the same cluster because they have the same entity types is not very interesting and does not help to create relation type clusters or labels that are descriptive in a way that is particularly useful e.g. for creating entity sketches. Furthermore, they compute external context based on surface order when the parse tree is available and grammatical relations could be used instead. They also use a phrase structure parser instead of a dependency parser, relying on constituency structure instead of taking advantage of grammatical dependencies such as deep subjects.

2.5 Summary

This chapter began by presenting the background to the relation extraction task. This was situated within the information extraction framework, which has the practical goal of extracting structured information from natural language text. This chapter also formulated a definition of a relation mention for the work in this thesis as a predicate ranging over two arguments, where an argument represents concepts, objects or peo-

ple in the real world and the relation predicate describes the type of stative association or interaction that holds between the things represented by the arguments.

Next, various approaches to domain adaptation were discussed. Generic techniques were motivated by two primary factors. First is the expense of supervised and partially supervised approaches. In the best case, they require only a small seed set of annotated data or a handful of examples of a given relation to move to a new domain. In the worst case, however, they require complete re-engineering including, for supervised machine learning, large-scale annotation and tuning of algorithm parameters. Second, supervised and partially supervised approaches require that the relation schema be known in advance. In reality, this is often not the case when moving to a new domain and certainly is not the case for ad-hoc applications like on-demand construction of relation tables from document collections. For example, Sekine (2006) describe such a system that summarises web search results. These limitations motivate generic relation extraction, which aims to devise models that learn the relation schema (the taxonomy of relation types to be identified) for a new domain as well instantiating an ontology with relation instances.

Various approaches to GRI and GRC were presented and several main shortcomings in the existing literature were identified:

- First, while previous approaches all focus on similarity modelling and clustering for GRE, there is a lack of standardised task definitions and evaluation frameworks. Differences in task definition are evident in the named entity recognition (NER) and coreference resolution (Coref) columns of Tables 2.2 and 2.4. With respect to GRC, differences are also evident in the instance level (Instance) column in Table 2.4. Differences in evaluation frameworks are evident in the data set (Data), gold standard reference information (Gold), entity pair sub-domain (Sub-Doms) and evaluation measure (Eval) columns in Tables 2.2 and 2.4. These differences make meaningful comparison across approaches impossible.

- Second, while some recent models from the literature incorporate constituent information from phrase structure parsers, they do not exploit governor-dependency information from dependency parsers. Additionally, previous models rely on direct matching of features for computing similarity, which fails to identify similarities between features with different surface strings but similar underlying (or latent) semantics. For example, the word tokens “hire” and “recruit” can both be used to describe a BUSINESS relation where a person works for an organisa-

tion, but this similarity would not be captured by the models from the existing literature.

- Third, while the claim that adaptation can be achieved without modification of model parameters is an important motivating factor, it has not been explicitly tested for GRE by evaluating model performance across different domains. Furthermore, the task has largely been developed and assessed without a concrete application in mind and, as a consequence, no real evaluation exists of the fully automatic end-to-end GRE task.

This thesis addresses these shortcomings by formalising the definition and evaluation of the generic relation extraction task, introducing state-of-the-art models based on dependency parsing and dimensionality reduction, explicitly evaluating the claim of modification-free domain adaptation and formalising an extrinsic evaluation based on extractive summarisation that serves as a test bed for end-to-end GRE.

Chapter 3

Task, Data and Evaluation

The generic relation extraction task is presented, which unifies the previously disjoint but closely related literatures on generic relation identification and generic relation characterisation. Data sets are described that are derived from publicly available corpora in the news and biomedical domains, allowing the claim of modification-free adaptation to be evaluated. Finally, evaluation measures are defined for the relation identification and characterisation tasks. The result is a rigorous and thorough framework for the evaluation of GRE across domains, including the introduction of statistical significance testing across entity pair sub-domains.

3.1 Introduction

As discussed in the previous chapter, the work in this thesis is motivated by a number of shortcomings in the literature including the lack of standardised task definitions and evaluation frameworks, which would allow meaningful comparison. This chapter moves towards the formalisation of the generic relation extraction (GRE) task. It proposes: 1) a combined framework for generic relation identification and characterisation, 2) standard data sets based on publicly available relation extraction data which allow evaluation of both generic relation identification and characterisation and 3) standard evaluation measures for generic relation identification and characterisation with respect to gold standard annotation. Section 3.2 contains a description of the combined framework for GRE. Section 3.3 contains a description of the data used here. Finally, Section 3.4 contains a description of evaluation measures, including a novel approach for the GRC task that is useful for error analysis.

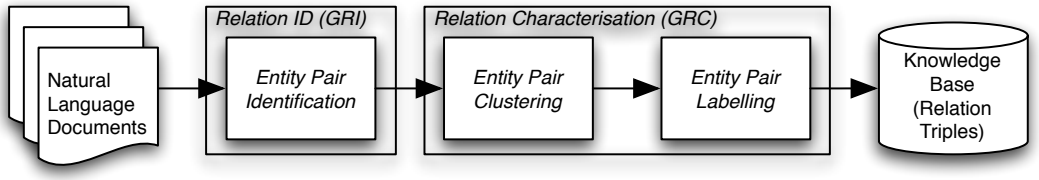


Figure 3.1: *Overview of main generic relation extraction sub-tasks. Relation identification comprises the sub-task of identifying relation-forming entity pairs. Relation characterisation comprises the sub-tasks of clustering entity pairs by relation type and choosing labels for clusters.*

3.2 The Generic Relation Extraction Task

Figure 3.1 gives an overview of the GRE task. Given input with pre-processing that includes entity mention markup and various linguistic annotations described below, the first step is generic relation identification (GRI), which identifies relation-forming entity pairs (defined here as relation mentions, consisting of pairs of entity mentions). The second step is generic relation characterisation (GRC), which is split into two sub-tasks. First, the GRC module induces a partition (or clustering) over the relation-forming entity mention pairs that groups them by relation type. Second, depending on the application, the GRC module annotates the identified clusters with automatically chosen labels that are descriptive of the relation type represented by the cluster.

Before proceeding here, it is useful to formalise some of the terminology that will be used in the rest of the thesis. Definitions for primary terminology can be found in Table 3.1. In addition, it is worth making a few usage notes for derivative terminology. The term *entity mention pair* can refer to any pair of entity mentions (whether they form a relation or not will depend on the context). The term *relation-forming entity mention pair* (also *co-occurring entity mention pair*) only refers to pairs that do form a relation. Finally, a *relation mention* can be untyped or typed depending on the context (generally untyped in the context of GRI where it is equivalent to a relation-forming entity mention pair and typed in the context of GRC).

In the rest of this section, the interfaces between each task or sub-task are described with reference to Figure 3.2, which contains example input and output derived from the gold standard data sets (described in Section 3.3).

Entity	A concept, object or person in the real world
Entity mention	An instance of a textual reference to an entity
Entity type	A label used to describe the taxonomic class of an entity
Entity schema	A taxonomy of possible entity types
Relation	A stative association that holds between two entities
Relation mention	An instance of a textual reference to a relation
Relation type	A label used to describe the taxonomic class of a relation
Relation schema	A taxonomy of possible relation types

Table 3.1: *Summary of terminology*

3.2.1 Example GRE System Input

The input to the GRE task consists of source documents with entity mention markup. Figure 3.2(a) contains several example sentences from the news and biomedical domains. The first column (S) contains the sentence identifier and the second column (Sentence Text) contains the text of the sentence (Note: Sentence 1 is from a broadcast news transcript that does not contain capitalisation). Entity mention boundaries are marked by square brackets with the type as a superscript on the opening bracket. In the third row, for example, the sentence ID is 3, the sentence text is "American saxophonist David Murray recruited Amidu Berry", and there are three entity mentions (i.e., the PLACE entity mention "American", the PERSON entity mention "David Murray" and the PERSON entity mention "Amidu Berry").

In addition to entity mention markup, the input documents contain linguistic information identified during pre-processing. Linguistic pre-processing includes sentence boundary identification, word tokenisation, part-of-speech tagging, identification of noninflected base word forms (lemmatisation), and dependency parsing. Preprocessing is described in detail in Section 3.3.

3.2.2 GRE Step 1: Generic Relation Identification

The first step in GRE is generic relation identification (GRI), where the goal is to identify relation-forming entity mention pairs. The input to the GRI sub-task is described above in Section 3.2.1 above and consists of sentences from source documents with entity mention markup. For the purpose of the intrinsic evaluation, gold standard entity mention annotation is used. While this does not reflect the accuracy of an end-to-end

a) Input to the generic relation extraction (GRE) task

S	Sentence Text
1	[PERSON martha stewart]'s company is registered as [ORGANISATION m.s. living omnimedia].
2	[PERSON Toefting] transferred to [ORGANISATION Bolton] from [ORGANISATION Hamburg].
3	[PLACE American] saxophonist [PERSON David Murray] recruited [PERSON Amidu Berry].
4	[*protein* Smooth muscle talin] prepared from chicken gizzard binds [*protein* skeletal muscle actin].
5	[*protein* Profilin] is believed to be an essential regulator of the [*source* actin cytoskeleton].
6	[*gene* Cdc3+] encodes [*protein* profilin], an [*protein* actin-monomer]-binding protein.

b) Output from the generic relation identification (GRI) module

R	S	Entity 1	Entity 2
1	1	"martha stewart"	"m.s. living omnimedia"
2	2	"Toefting"	"Bolton"
3	2	"Toefting"	"Hamburg"
4	3	"David Murry"	"American"
5	3	"David Murry"	"Amidu Berry"
6	4	"Smooth muscle talin"	"skeletal muscle actin"
7	5	"Profilin"	"actin cytoskeleton"
8	6	"Cdc3+"	"profilin"
9	6	"profilin"	"actin-monomer"

c) Output from the generic relation characterisation (GRC) module

R	S	Entity 1	Entity 2	C	Cluster Label
1	1	"martha stewart"	"m.s. living omnimedia"	1	EMPLOY-EXECUTIVE
2	2	"Toefting"	"Bolton"	2	SPORTS-AFFILIATION
3	2	"Toefting"	"Hamburg"	2	SPORTS-AFFILIATION
4	3	"David Murry"	"American"	3	CITIZEN-OR-RESIDENT
5	3	"David Murry"	"Amidu Berry"	4	BUSINESS
6	4	"Smooth muscle talin"	"skeletal muscle actin"	5	BIND
7	5	"Profilin"	"actin cytoskeleton"	6	CONTROL
8	6	"Cdc3+"	"profilin"	7	ENCODE
9	6	"profilin"	"actin-monomer"	5	BIND

Figure 3.2: *Example input and output for GRE modules.*

system, it isolates the errors that are due to the GRI module. The accuracy of the fully automatic system is measured in the extrinsic evaluation in Chapter 6 and demonstrated in the example GRE output in Section 3.2.4 below.[1] Here, all pairs of entity mentions that occur in the same sentence are considered to be candidate relation mentions. Only considering intra-sentential relation mentions is a simplifying assumption. However, in the three data sets used for the current work (which all contain at least 900 gold standard relation mentions), there is only a one instance of a gold standard relation mention where the entity mentions are in different sentences (see Section 3.3 for details).

The GRI task, therefore, is to consider each pair of entity mentions within a sentence and determine whether the pair constitutes a relation mention or not. The output from GRI is illustrated in Figure 3.2(b). The first column (R) contains a relation mention identifier. The second column (S) is the sentence identifier, which links the relation-forming pairs back to the source sentences in Figure 3.2(a). And, the third (Entity 1) and fourth (Entity 2) columns contain the individual entity mentions that make up the relation-forming pair. The fifth row, for example, states that there is a relation mention (with identifier 5) in Sentence 3 between the two PERSON entity mentions "David Murray" and "Amidu Berry". At this point, there is no information about the type of relation, only about the existence of some relation mention between two entity mentions with unspecified type.

3.2.3 GRE Step 2: Generic Relation Characterisation

The second step in GRE is generic relation characterisation (GRC), where the goal is to annotate each relation mention with a label that describes the relation type. The input to the GRC sub-task is the output from the GRI sub-task described above in Section 3.2.2 and consists of sentences from the source document with entity mentions and relation-forming pairs identified. For the purpose of the intrinsic evaluation here, gold standard entity and relation-forming entity pair annotations are used. While this does not reflect the accuracy of an end-to-end system, it isolates the errors that are due to

[1]The choice to focus on the relation extraction problem is also justified by the fact that NER is a relatively well understood task. F-scores on the newswire and broadcast news data from ACE 2005 range from 0.72 to 0.77 for the top systems (NIST, 2006). On other newswire data sets, f-scores are near 0.90 (e.g., Sang and Meulder, 2003). On biomedical data sets, results are similar to ACE 2005, with Alex et al. (2007) reporting results of 0.71 on a protein-protein interaction data set. There are various free and commercial off-the-shelf systems for NER (e.g., Bikel et al., 1999; Curran and Clark, 2003; Alias-i, 2007). Furthermore, a number of authors have demonstrated effective bootstrapping in various domains (e.g., Collins and Singer, 1999; Thompson et al., 1999; Jones et al., 2003; Hachey et al., 2005; Vlachos and Gasperin, 2006), suggesting that adaptation approaches for NER are reasonably mature.

the GRC module. As mentioned above, the accuracy of the fully automatic system is measured in the extrinsic evaluation in Chapter 6 and demonstrated in the example GRE output in Section 3.2.4 below. The GRC sub-task proceeds in two main steps corresponding to the fifth (C) and sixth (Cluster Label) columns of Figure 3.2(c).

First, the system induces a partition (or clustering) over the relation-forming pairs, where the goal of the clustering is to group them by relation type. For the current work, each relation mention (i.e., pair of co-occurring entity mentions) from the output of the GRI sub-task is an instance for clustering. In other words, the clustering instance level is entity pair tokens instead of entity pair types (see Section 2.4.2.3 for further discussion). This choice is based on the argument from Zhang et al. (2005) that different types of relations can exist between different mentions of the same two underlying entities and is also motivated by the fact that this allows direct linking from GRC output back to individual relation mentions, which is useful for error analysis and data exploration. The clustering is based on features of the sentential context of the relation mention, e.g. the underlined words in the following text snippets:

> "[^{protein} Smooth muscle talin] prepared from chicken gizzard binds [^{protein} skeletal muscle actin]"
>
> "[^{protein} profilin], an [^{protein} actin-monomer]-binding protein"

The output of the clustering task is illustrated in the fifth column (C) of Figure 3.2(c), where the number corresponds to the cluster identifier to which the relation-forming pair has been assigned. So, for example, Relation Mentions 6 and 9 are deemed to have the same relation type and therefore they are placed together in Cluster 5.

Finally, the system identifies a label for each cluster that is descriptive of the relation type represented by the cluster. In the example above, the label for Cluster 5 is BIND. It is also the only word that shows up in the context of both Relation Mention 6 and Relation Mention 9 as indicated by the double underline in the example text snippets above.

3.2.4 Example GRE System Output: Entity Sketches

One possible application of GRE is as a general purpose tool for creating entity sketches for any document set where entities can be identified. Figure 3.3 contains example entity sketches for three entities in the news domain[2] using fully automatic GRE (in-

[2]The system is run over the DUC development data, which is described in detail in Chapter 6. This consists of small document collections on specific topics, which is the reason for the thematic similarity

cluding NER, relation identification, model order selection, clustering and cluster labelling). Model order selection and cluster labelling use the approaches from Chen et al. (2005) (see Section 2.4.2.2 of the previous chapter). Figure 3.3(a) contains a sketch for the PERSON entity "Neil Bush"; Figure 3.3(b) for the ORGANISATION entity "NRA"; Figure 3.3(c) for the PERSON entity "Boesky". The first column (R) of each sketch contains an identifier for the entity pair. For each pair, the second entity (ENTITY) and the cluster labels (LABELS) are listed. Labels are presented in rank order and a manually assigned cluster name is given in parentheses at the end of the label list. Example text snippets (TEXT1, TEXT2) containing individual mentions of the given relation are also given to illustrate the source data. These are chosen based on how representative the pair is of the cluster to which it belongs.[3]

In Figure 3.3(b), for example, Relation 1 for the ORGANISATION entity "NRA" (i.e., the National Rifle Association, a personal firearm advocacy group) is related to the LOCATION entity "Washington". The pair is grouped into a cluster that can be interpreted as representing LOCATED relations. The first label for this cluster is "r_in", where "r" indicates that the label is from the dependency path and "in" indicates that the dependency type is the preposition in. The labels also include "w_office", where "w" indicates that the label is from the words in the context of the entity pair and "office" is the word. The first example text snippet (TEXT1) is "he had an aide call [ORGANISATION NRA] officials in [LOCATION Washington]" and the second example text snippet (TEXT EX2) is "[ORGANISATION NRA] headquarters in [LOCATION Washington]". The intuition behind this approach to creating entity sketches is that an entity can be described by the relations that it takes part in. The three relations for "NRA" indicate that 1) the NRA has headquarters located in Washington, 2) the NRA has a business or membership relationship with Dennis DeConcini (a U.S. Senator from Arizona), and 3) James Jay Baker is employed as a lobbyist for the NRA.

between the relations (e.g., both relations for "Neil Bush" in Figure 3.3(a) are from a topic that is concerned with involvement in the financial scandals at the American Savings and Loan institutions in the 1980s and 1990s).

[3]This is done by selecting sentences containing pairs with the minimum distance from the mean (centroid) feature vector for the cluster to which the pair belongs. Distance is measured over LDA topic vectors using Kullback-Leibler divergence. LDA and Kullback-Leibler divergence are described in Chapter 5.

a) Entity sketch for PERSON entity "Neil Bush"

R	Relation	Description
1	ENTITY:	"Denver" (LOCATION)
	LABELS:	r_nn, r_subj, r_of, r_conj, r_person, r_lex-mod, r_appo, r_obj, r_gen, r_inside, w_said, ... (LOCATED)
	TEXT1:	"[PERSON Neil Bush], a [LOCATION Denver] oilman"
	TEXT2:	"in [LOCATION Denver], [PERSON Neil Bush] became"
2	ENTITY:	"Amoco" (ORGANISATION)
	LABELS:	r_in, r_conj, r_gen, r_appo, r_at, r_mod, w_office, r_for, r_lex-mod, r_fc, r_to, ... (EMPLOYMENT)
	TEXT1:	"[PERSON Neil Bush] worked as a negotiator for [ORGANISATION Amoco]"
	TEXT2:	"[PERSON Neil Bush] was hired as a landman by [ORGANISATION Amoco]"

b) Entity sketch for ORGANISATION entity "NRA"

R	Relation	Description
1	ENTITY:	"Washington" (LOCATION)
	LABELS:	r_in, r_conj, r_gen, r_appo, r_at, r_mod, w_office, r_for, r_subj, r_nn, r_lex-mod, ... (LOCATED)
	TEXT1:	"he had an aide call [ORGANISATIONNRA] officials in [LOCATION Washington]"
	TEXT2:	"[ORGANISATIONNRA] headquarters in [LOCATION Washington]"
2	ENTITY:	"DeConcini" (PERSON)
	LABELS:	r_conj, w_secretary, r_fc, r_of, r_gen, r_appo, w_president, w_said, r_subj, w_chairman, r_obj, ... (MEMBERSHIP)
	TEXT1:	"[PERSON DeConcini] accused the [ORGANISATION NRA] of lies"
	TEXT2:	"[PERSON DeConcini], an [ORGANISATION NRA] 'Person of the Month"'
3	ENTITY:	"James Jay Baker" (PERSON)
	LABELS:	r_conj, r_of, r_gen, w_secretary, r_appo, w_chairman, r_obj, r_lex-mod, r_subj, r_person, r_in, ... (EMPLOYMENT)
	TEXT1:	"[PERSON James Jay Baker], the [ORGANISATION NRA]'s top lobbyist"
	TEXT2:	"[PERSON James Jay Baker], the [ORGANISATION NRA]'s chief lobbyist."

c) Entity sketch for PERSON entity "Boesky"

R	Relation	Description
1	ENTITY:	"Milken" (PERSON)
	LABELS:	r_conj, r_person, r_nn, r_appo, r_lex-mod, r_of, w_president, r_subclass, r_mod, r_subj, w_judge, ... (BUSINESS)
	TEXT1:	"[PERSON Milken] and [PERSON Boesky] ended up striking deals"
	TEXT2:	"[PERSON Milken] used [PERSON Boesky]'s firm to hide illegal stock trading"

Figure 3.3: *Example output: GRE for entity sketches.*

3.3 Data

Two corpora are adapted for tuning and evaluation of systems addressing the GRE task, allowing for comparative evaluation across the news and biomedical domains. For the news domain, we use the data from the Automated Content Extraction shared tasks.[4] For the biomedical domain, we use the BioInfer data.[5]

In order to compare results on these two corpora, they are converted to a standard format following the three basic steps in Figure 3.4. Step 1 takes the raw corpora as input and outputs a standard XML format for RE data. Core output is an XML document containing sentence and word token markup with entity and relation mentions specified using token standoff (see Appendix A). In Step 2, linguistic information is added into the XML format. This pre-processing includes part-of-speech tagging and lemmatisation using the LT-TTT tools (Grover et al., 2000). LT-TTT is a general purpose text tokenisation tool. It is implemented using LT-XML2, a collection of generic XML text processing tools (Grover et al., 2006). The version used here is an early version of the upcoming LT-TTT2 release. Lemmatisation in LT-TTT is performed using the Morpha tool (Minnen et al., 2000). Step 2 also includes dependency parsing using Minipar (Lin, 1998). (See Chapter 4 for further details of Minipar output and Appendix A for details of the dependency markup in the XML documents.) Finally, in Step 3, the data is normalised such that relation mentions are between named entity mentions where possible. For the ACE data, this consists primarily of a mapping of a number of nominal entity mentions (e.g., "one half of PBS") to named entity mentions (e.g., "Amidu Berry"). For the BioInfer data, this consists primarily of a mapping from n-ary to binary relation mentions. Details of the respective transformations are given in the following sections (3.3.1 and 3.3.2).[6]

Furthermore, for the GRE evaluation here, relation mentions are required to be between exactly two entity mentions that are in the same sentence[7] and are distinct siblings. First, the requirement that the entity mentions be *distinct* removes reflexives, which are relation mentions where either both entity mentions are identical or the type

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

[5] http://mars.cs.utu.fi/BioInfer/

[6] I am waiting to hear from the Linguistic Data Consortium about re-distribution of the modified ACE data. The modified version of the BioInfer data will be made available free of charge under the same license terms as the original BioInfer data set.

[7] There are seven relation mentions in ACE 2004 that cross sentence boundaries. However, all of them are due to errors in the automatic boundary identification. In ACE 2005, there are six cross-sentence relation mentions, five of which are due to sentence boundary errors. In the BioInfer data, there are no relation mentions that cross sentence boundaries because annotation is at the sentence level.

1	Re-factoring:	Convert data to REXML format
2	Pre-processing:	Add linguistic markup
3	Re-annotation:	Normalise annotation for GRE

Figure 3.4: *Basic steps for standardising RE corpora to allow comparative evaluation.*

and normalised surface strings for both entity mentions are identical. Reflexive relation mentions are sometimes introduced erroneously from the annotation, e.g. the SUBSIDIARY("afghanistan","afghanistan") relation mention in "Afghanistan's post-Taliban government". The original relation mention in ACE is SUBSIDIARY("government","afghanistan"). However, because "afghanistan" and "government" are annotated, rather strangely, as being coreferent, Mapping Rule 4 (described in Section 3.3.1) fires and the relation mention ends up being SUBSIDIARY("afghanistan","afghanistan"). Second, the requirement that the entity mentions be *siblings* removes relation mentions where the entity mentions are not immediately contained within the same embedding entity mention or sentence. The primary effect here is that pairs where one entity mention is embedded within the other (i.e., one is a parent or grandparent of the other) are not considered. This also means that other long distance relationships within the entity mention constituent tree (e.g., cousins) are not considered. For example, in the text snippet "E-cadherin/plakoglobin complexes", the CHANGE/PHYSICAL("E-cadherin", "plakoglobin") relation mention is kept, but the following two relation mentions are ignored:

OBJECT-COMPONENT("E-cadherin/plakoglobin complexes", "E-cadherin")
OBJECT-COMPONENT("E-cadherin/plakoglobin complexes", "plakoglobin")

Finally, seven entity pair subsets are chosen for each data set based on several criteria. Relation types are considered to be outliers and filtered if they have less than 3 total mentions. Also, entity pair domains are only used for GRC if they have 30 or more total mentions and 2 or more distinct relation types. Table 3.2 contains overview information about the resulting data sets for generic relation identification (GRI) and characterisation (GRC). For the GRI data, the Total Instances row contains the count of all possible pairs of entity mentions occurring in the same sentence. And the Proportion True Instances rows contain the percentage of true relation mentions according to the gold standard with respect to the total number of possible entity mention pairs. For the GRC data, the Total Instances row contains the total number of true relation mentions

	Development (ACE 2004)	News Test (ACE 2005)	Biomedical (BioInfer)
GRI DATA			
Number Ent Pair Subsets	7	7	7
Total Instances	9253	3012	5843
Proportion True Instances (Micro Ave)	11.4%	9.1%	27.2%
Proportion True Instances (Macro Ave)	13.1%	11.0%	26.8%
GRC DATA			
Number Ent Pair Subsets	7	7	7
Total Instances	1400	877	1301
Total Relation Types	14	15	4
Types Per Ent Pair Subset (Macro Ave)	3.9	3.3	2.9

Table 3.2: *Summary information for GRE data sets.*

according to the gold standard. The Total Relation Types row contains the number of relation types in the full data set (i.e., including all entity pair subsets). And the Types Per Ent Pair Subset row contains the mean number of relation types across entity pair subsets.

3.3.1 News IE Data: ACE

3.3.1.1 Overview

The data for the news domain is derived from the IE corpora that were prepared for the Automatic Content Extraction (ACE) shared tasks.[8] For the experiments in this thesis, the data from the 2004 evaluation is used for development and the data from the 2005 evaluation is used as a held-out test set. Only newswire and broadcast news materials are used. The 2004 (development) and 2005 (news test) data are non-overlapping. Data from the ACE evaluations was also used by Chen et al. (2005, 2006). Using data with unbiased gold standard relation annotation allows rapid, automatic evaluation schemes, which means results can be reimplemented and compared.

[8]The US National Institute for Standards and Technology (NIST) sponsors ACE and releases the training data through the Linguistic Data Consortium (LDC). The NIST project page for ACE is at `http://www.nist.gov/speech/tests/ace/`. The LDC project page for ACE is at `http://projects.ldc.upenn.edu/ace/`.

Source	Type	Epoch	Num Docs	
DEVELOPMENT (ACE 2004)				
Associated Press	Newswire	2000/10-12	73	(21.0%)
Cable News Network	Broadcast News	2000/10-12	63	(18.1%)
Voice of America	Broadcast News	2000/10-12	57	(16.5%)
New York Times	Newswire	2000/10-12	55	(15.8%)
Public Radio International	Broadcast News	2000/10-12	38	(10.9%)
American Broadcasting Company	Broadcast News	2000/10-12	25	(7.2%)
MSNBC	Broadcast News	2000/10-12	19	(5.5%)
National Broadcasting Company	Broadcast News	2000/10-12	18	(5.2%)
NEWS TEST (ACE 2005)				
Cable News Network	Broadcast	2003/03-06	177	(59.4%)
CNN Headline News	Broadcast	2003/03-06	40	(13.4%)
Associated Press	Newswire	2003/03-06	38	(12.8%)
Agence France Presse	Newswire	2003/03-06	27	(9.1%)
Xinhua News Agency	Newswire	2003/03-06	13	(4.4%)
New York Times	Newswire	2003/03-06	3	(1.0%)

Table 3.3: *Sources for ACE 2004 and 2005 news data.*

The final set of documents used here is summarised in Table 3.3. The first column (Source) corresponds to the name of the organisation from which the data was obtained. The second column (Type) corresponds to the media type of the data source. Newswire indicates that the data was obtained from a printed news feed. Broadcast News indicates that the data is obtained from the transcript of a spoken news programme. The data from broadcast news sources is generally well edited, though does not contain capitalisation. For the purposes of evaluating relation identification and characterisation given gold standard entities, the newswire and broadcast news data are similar enough to be combined. The third and fourth columns correspond to the range of months during which the sources were published (Epoch) and the number and distribution of documents from each source organisation (Num Docs). The total number of documents is 348 and 298 respectively for the development and news test data. The overall newswire-broadcast news splits are approximately 36.8%-63.2% and 27.2%-72.8%.

3.3.1.2 Re-Annotation

After converting the ACE data to the REXML format and performing pre-processing, the annotation is normalised for the GRE task. This is motivated primarily by the complex, nested nature of the deep linguistic entity annotation in ACE (see description of mapping rules below). This is also motivated by the desire to create a simplified entity scheme that allows statistical significance testing across entity pair sub-domains (as described in Section 3.4.4). Re-annotation proceeds in three steps: 1) mapping nominal entity mentions to named entity mentions, 2) filtering relation and entity mentions that are not relevant to the evaluation of the GRE task, and 3) converting entity and relation types to the final schema.

The mapping performed in the first step is motivated by the prevalence of nominal entity mentions in ACE, where entities can be referenced by their name (i.e., named mention), by a common noun or noun phrase (i.e., nominal mention) or by a pronoun (i.e., pronominal mention). The mapping is also facilitated by the presence of detailed linguistic annotation which makes it possible to automatically map many nominal entity mentions to named entity mentions. Several aspects of the detailed ACE annotation are used in the mapping rules: entity extent, entity type and entity mention type. In the following description of the mapping rules, this information is represented by typesetting conventions illustrated in the following text snippet:

"$[^{per}_{nam}$ Amidu Barry], $[^{per}_{nom}$ one half of $[^{org}_{nam}$ PBS]]".

where the boundaries of the full entity mention extent are indicated by square brackets, entity type is a superscript on the opening square bracket, and entity mention type is a subscript on the opening square bracket. There are three entity mentions, including one nominal mention and two named mentions:

1. "Amidu Barry" with type PERSON and mention type *named* (NAM)
2. "one half of PBS" with type PERSON (PER) and mention type *nominal* (NOM)
3. "PBS" with type ORGANISATION (ORG) and mention type *named*

Table 3.4 contains a list of possible entity mention types (Label) with a short description (Description), an example (Example) and the number and proportion of occurrences in the ACE 2004 and the ACE 2005 data sets.[9] ACE 2004 entity types include

[9]The rules used here only consider nominal and prenominal entity mentions for possible mapping. It may also be possible to map from the other, less frequent nominal mention types. However, the

Label	Description	Example	ACE 2004		ACE 2005	
NAM	Named entity reference	"John", "Fargo"	6903	(30.4%)	4586	(25.4%)
PRO	Pronominal reference	"they", "her"	5119	(22.5%)	4684	(25.9%)
NOM	Nominal reference	"the lawyer"	4853	(21.3%)	4001	(22.1%)
PRE	Prenominal reference	"[Labour] nominee"	2992	(13.2%)	2489	(13.8%)
BAR	Unquantified nominals	"lawyers"	1990	(8.8%)	1673	(9.3%)
WHQ	WH words and specifiers	"UK, [where] . . . "	511	(2.2%)	367	(2.0%)
HLS	Headless mentions	"the biggest"	194	(0.9%)	152	(0.8%)
PTV	Partitive constructions	"some of us"	111	(0.5%)	134	(0.7%)
MWH	Multiple-word heads	"20 men and women"	63	(0.3%)	0	(0.0%)

Table 3.4: *Entity mention types in the ACE source data. Columns contain the mention type label (Label), a description (Description), an example (Example) and the count and percentage of occurrences for ACE 2004 and ACE 2005.*

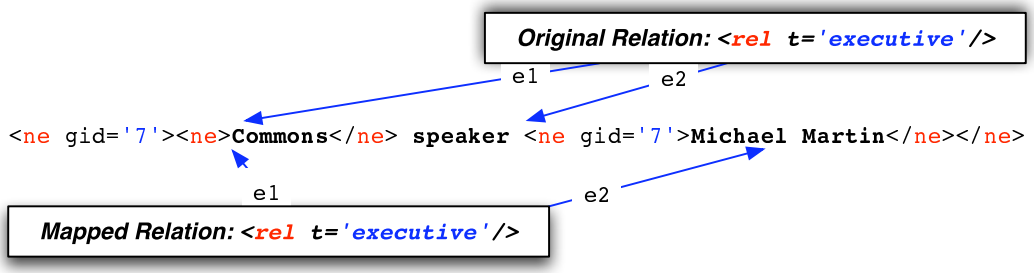
PERSON (PER), ORGANISATION (ORG), FACILITY (FAC), LOCATION (LOC), GEOGRAPHICAL/POLITICAL (GPE), VEHICLE (VEH) (LDC, 2004b). ACE 2005 entity types include PERSON (PER), ORGANISATION (ORG), GEOGRAPHICAL/SOCIAL/POLITICAL (GPE), LOCATION (LOC), FACILITY (FAC), VEHICLE (VEH) and WEAPON (WEA) (LDC, 2005a).

Furthermore, entity heads (see Section 2.4.2.3 for a description of what a head is) are annotated in ACE and indicated here by an underscore such as "one" in "one half of PBS". Finally, entity mention coreference markup in ACE is central to the re-annotation as mappings are only allowed between two mentions that refer to the same underlying entity (e.g., "one half of PBS" and "Amidu Barry" above). In the following description, coreference will either be noted or obvious from the context. In addition, some of the mapping rules described below make use of aspects of the linguistic pre-processing introduced at the beginning of Section 3.3 above.

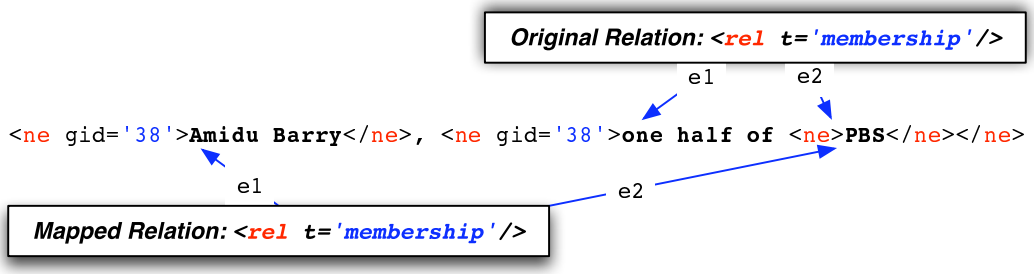
Figure 3.5 contains three example mappings from different rules. The first mapping is possible because the entity mention "Michael Martin" is coreferent with and embedded within the entity mention "Commons speaker Michael Martin" and because the latter is annotated as having a prenominal mention type, indicating that it occurs in a modifying position before another noun. Thus, the embedded EXEC-

feasibility would have to be carefully investigated to ensure the mappings are sensible. Some mention types (e.g., unquantified nominals) would probably be safe, but others (e.g., headless mentions) would require more careful consideration.

a) Mapping Rule 1: Prenominal ↦ Embedded Coreferent



b) Mapping Rule 5: Nominal ↦ Left Adjacent Coreferent



c) Mapping Rule 6: Nominal ↦ Right Adjacent Coreferent

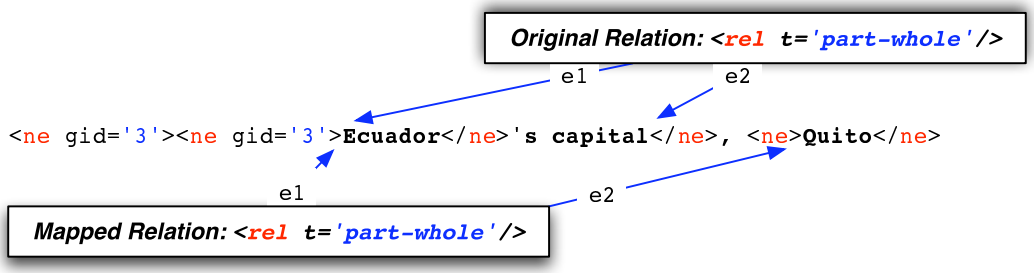


Figure 3.5: *Example ACE mappings from nominal to named entity mentions.*

UTIVE("Commons", "Commons speaker Michael Martin") relation mention is converted to EXECUTIVE("Commons", "Michael Martin").

Mapping Rule 5 is possible because the entity mention "Amidu Barry" is coreferent with and immediately to the left of the entity mention "one half of PBS" and because the latter is annotated as having a nominal mention type. Thus, the embedded relation mention MEMBERSHIP("one half of PBS","PBS") with nominal entity mention "one half of PBS" is converted to the non-embedded, fully named relation mention MEMBERSHIP("Amidu Barry","PBS"). Mapping Rule 6 is analogous except that it maps to a named entity mention to the right, converting the embedded relation mention PART-WHOLE("Ecuador","Ecuador's capital") with nominal entity mention "Ecuador's capital" to the entity mention PART-WHOLE("Ecuador","Quito").

The full list of mapping rules is found in Table 3.5. The first column (#) lists the rule number. The second column (Description of Mapping Rule) contains a brief textual description of the mapping rule, where the mention type of the original entity mention is on the left, followed by the $\mapsto$ symbol, followed by a specification of how the target entity mention for the mapping rule is identified. Finally, the third (ACE 2004) and fourth (ACE 2005) columns contain a count of how many times each rule fired and the percentage of total firings accounted for by each rule for the respective data sets. Rules are ordered from those that are the most constrained to those that are the least constrained. In developing these rules, no errors were identified in the mapped relation mentions from Rules 1 through 12. Mapping errors resulting from Rules 13 and 14 are discussed below.

Rule 1 is illustrated in Figure 3.5 and described above. It is the only rule that maps from a prenominal entity mention. The remaining rules all map from a nominal entity mentions to a named or pronominal entity mention.

Rule 2 maps to a coreferent and embedded entity mention occurring immediately to the left of the head of the original entity mention, e.g.:

> "[$^{gpl}_{nom}$ [$^{gpl}_{nam}$ Indonesia]'s war-torn [$^{gpl}_{nam}$ Aceh] province]"
>
> PART-WHOLE("Indonesia's war-torn Aceh province", "Indonesia")
> PART-WHOLE("Aceh", "Indonesia").

Rule 3 maps to any coreferent and embedded entity mention, e.g.:

> "[$^{gpl}_{nom}$ [$^{per}_{nam}$ gore]'s home state of [$^{gpl}_{nam}$ tennessee]]"
>
> CITIZEN-OR-RESIDENT("gore", "gore's home state of tennessee")
> CITIZEN-OR-RESIDENT("gore", "tennessee").

#	Description of Mapping Rule			ACE 2004		ACE 2005	
1	Prenom.	$\mapsto$	Embedding Coreferent	191	(26.6%)	104	(20.6%)
2	Nominal	$\mapsto$	Embedded Left Adjacent Prenom.	30	(4.2%)	34	(6.7%)
3	Nominal	$\mapsto$	Embedded Coreferent	41	(5.7%)	73	(14.5%)
4	Nominal	$\mapsto$	Embedding Coreferent	11	(1.5%)	3	(0.6%)
5	Nominal	$\mapsto$	Left Adjacent Coreferent	178	(24.8%)	69	(13.7%)
6	Nominal	$\mapsto$	Right Adjacent Coreferent	133	(18.5%)	106	(21.0%)
7	Nominal	$\mapsto$	Left Adjacent Coreferent, Skip Copula	35	(4.9%)	31	(6.2%)
8	Nominal	$\mapsto$	Right Adjacent Coreferent, Skip Copula	3	(0.4%)	1	(0.2%)
9	Nominal	$\mapsto$	Left Adjacent Coreferent, Skip Verb+TO_BE	1	(0.1%)	1	(0.2%)
10	Nominal	$\mapsto$	Right Adjacent Coreferent, Skip Verb+TO_BE	0	(0.0%)	0	(0.0%)
11	Nominal	$\mapsto$	Left Adjacent Coreferent, Skip Cop and Coreferring Ents	3	(0.4%)	1	(0.2%)
12	Nominal	$\mapsto$	Right Adjacent Coreferent, Skip Cop and Coreferring Ents	1	(0.1%)	0	(0.0%)
13	Nominal	$\mapsto$	Left Coreferent	89	(12.4%)	73	(14.5%)
14	Nominal	$\mapsto$	Right Coreferent	3	(0.4%)	8	(1.6%)

Table 3.5: *Full list of rules for mapping from nominal to named entity mentions in ACE. Columns contain the rule identifier (#), the rule description (Description), and the number and percent of firings for ACE 2004 and ACE 2005.*

Rule 4 maps to any coreferent and embedding entity mention, e.g.:

> "[$^{gpl}_{nam}$ [$^{gpl}_{nom}$ the [$^{gpl}_{nom}$ West African] nation] of Senegal]"
>
> PART-WHOLE("the West African nation", "West African")
> PART-WHOLE("Senegal", "West African")

The mapping from the nominal entity mention "the West African nation" to the named entity mention "Senegal" is possible because the full extent of the target entity mention is "the West African nation of Senegal", which is subsequently shortened to "Senegal" by keeping only the string annotated as the entity mention head.

Rules 5 and 6 map to immediately adjacent coreferent entity mentions. These rules are illustrated in Figure 3.5 and described above. Rules 7 and 8 map to coreferent entity mentions found respectively to the left or to the right, that have only a copular verb phrase (i.e., a verb phrase where the lemma of the main verb is "be") and any number of adverbs intervening, e.g.:

> "[$^{per}_{nom}$ The last [$^{gpl}_{nam}$ U.S.] president to visit [$^{gpl}_{nam}$ Vietnam]] was [$^{per}_{nam}$ Nixon]"
>
> EMPLOY-EXECUTIVE("The last U.S. president to visit Vietnam", "U.S.")
> EMPLOY-EXECUTIVE("Nixon", "U.S.").

Rules 9 and 10 map again to coreferent entity mentions found respectively to the left or to the right. However, they allow two intervening verb phrases so long as the second is the infinitival copular (i.e., "to be") and any number of adverbs, e.g.:

> "[$^{per}_{nam}$ Bush] is probably going to be [$^{per}_{nom}$ the next [$^{gpl}_{nam}$ U.S.] president]"
>
> EMPLOY-EXECUTIVE("the next U.S. president", "U.S.")
> EMPLOY-EXECUTIVE("Bush", "U.S.").

Rules 11 and 12 map once again to coreferent entity mentions found respectively to the left or to the right. In this instance, however, they allow any number of coreferent entity mentions to intervene between the original nominal entity mention and the target named or pronominal entity mention, e.g.:

> "[$^{per}_{nam}$ Card] is [$^{per}_{nom}$ a [$^{gpl}_{nam}$ Washington] insider] and [$^{per}_{nom}$ a lobbyist for [$^{org}_{nam}$ General Motors]]"
>
> EMPLOY-STAFF("a lobbyist for General Motors", "General Motors")
> EMPLOY-STAFF("Card", "General Motors")

where both nominal entity mentions (i.e., "a Washington insider" and "a corporate lobbyist for the automobile industry and General Motors") are coreferent with the named entity mention "Card".

Finally, Rules 13 and 14 are general rules that map to any coreferent entity mentions found respectively to the left or to the right, e.g.:

> "[$^{org}_{nom}$ [$^{per}_{nam}$ martha stewart]'s company], officially known as [$^{org}_{nam}$ m. s. living omnimedia]"
>
> EMPLOY-EXECUTIVE("martha stewart", "martha stewart's company")
> EMPLOY-EXECUTIVE("martha stewart", "m. s. living omnimedia").

The extremely general nature of Rules 13 and 14 results in some questionable mappings. For the most part, the questionable relation mentions resulting from this mapping were not clearly erroneous but rather unrealistic in the sense that any tractable GRE system would be unlikely to be able to recover some of the resulting long distance relation mentions. For example, Rule 14 triggers the following mapping:

> "[$^{per}_{nam}$ Chapman]'s pursuit of publicity occupied much of the discussion, despite [$^{per}_{pro}$ his] professed wish to return to the anonymity that had plagued [$^{per}_{pro}$ him], as [$^{per}_{nom}$ a security guard in [$^{gpl}_{nam}$ Hawaii]"
>
> CITIZEN-OR-RESIDENT("security guard in Hawaii", "Hawaii")
> CITIZEN-OR-RESIDENT("Chapman", "Hawaii").

However, it would be more unrealistic to keep the embedded annotations, which are not representative of the relation mentions that would be discovered by the GRE task.

To get a sense for the accuracy of Rules 13 and 14, a random sample of twenty firings were inspected. Among these, four (20.0%) create relation mentions that are questionable or that could arguably have been mapped to a more suitable target entity mention. E.g., Rule 13 triggers the following mapping:

> "[$^{per}_{nam}$ Ehud Barak] won the endorsement of [$^{org}_{nom}$ [$^{per}_{pro}$ his] Labor party] as [$^{per}_{nom}$ [$^{org}_{pro}$ it]'s candidate for Prime Minister]"
>
> MEMBER-OF-GROUP("it's candidate for Prime Minister", "it")
> MEMBER-OF-GROUP("Ehud Barak", "it")

Here, the pronominal entity mention "it" could arguably be mapped to the named entity mention "Labor". This could be addressed by extending the current rules to map from pronominal to named entity mentions where this is possible. However, the knock-on effects would have to be carefully investigated. Furthermore, in this instance, the annotators actually failed to mark Labor as a named entity mention. In the current work, the mapping is allowed to fire and the resulting relation mentions are kept in the final data.

A similar problem is encountered when a nominal entity mention is mapped out of an embedded entity mention, but the other entity mention is a possessive prenominal pronoun that is not mapped, e.g.:

> "[$^{per}_{nom}$ [$^{per}_{nam}$ Gore]'s press secretary], [$^{per}_{nam}$ Chris Lehane], made it clear in an interview that [$^{per}_{nom}$ [$^{per}_{nam}$ Gore] aides] do not feel bound by [$^{per}_{nom}$ [$^{per}_{pro}$ their] candidate]'s pledge."
>
> BUSINESS("their", "their candidate")
> BUSINESS("their", "Gore")

This could be addressed by not using relation mentions with possessive pronominal entity mentions or only using them when the possessive pronoun immediately precedes the second entity mention in the relation and the second entity mention is a named reference. Again, the knock-on effects would have to be carefully investigated. This questionable mapping occurred 2 times in the sample. In the current work, the mapping is allowed to fire and the resulting relation mentions are kept in the final data.

The fourth problem with Rules 13 and 14 from the sample is due to oddities in the annotation guidelines or execution, e.g.:

> "The summit , which is being sponsored by [$^{gpl}_{nam}$ the European Union], is meant to show [$^{gpl}_{nom}$ the [$^{gpl}_{nam}$ Balkan] states] that [$^{gpl}_{nam}$ the EU] is preparing to welcome [$^{gpl}_{pro}$ them] into [$^{gpl}_{nom}$ the [$^{gpl}_{pre}$ European] family]."
>
> GPE-AFF-OTHER("the European family", "European")
> GPE-AFF-OTHER("the European Union", "European")

Here, the difficulty is in the complex semantics of the entity mention "the European family". This can be understood to refer to the member countries of the European Union at the time. This suggests that it should actually be annotated as being coreferent with the other entity mentions referring to the European Union, which it is not. In the current work, the mapping is allowed to fire and the resulting relation mentions are kept in the final data.

In the next step in the re-annotation process, certain entity and relation mentions are filtered. The motivation here is again to simplify the deep, linguistic level of annotation in the ACE data. First, all entity mentions that do not have a mention type of named (NAM), pronominal (PRO) or prenominal (PRE) are filtered. (The full list of entity mention types can be viewed by referring back to Table 3.4.) This serves to remove all nominal mentions, which are not recognised by most NER systems. Prenominal mentions are kept because they are often names (e.g., "Labour" in "[$^{per}_{nom}$ [$^{org}_{pre}$ Labour] nominee]"), though not always (e.g., "British prime minister" in "[$^{per}_{nam}$ [$^{per}_{pre}$ [$^{gpl}_{pre}$ British] prime minister] Tony Blair]"). The ACE 2005 data actually distinguishes between named and non-named prenominal mentions. However, the ACE 2004 (development) data does not, so the distinction is ignored for the evaluation here.

Label	Description	Example	ACE 2004		ACE 2005	
SPC	Specific referential	"a drop", "Perth"	18356	(80.7%)	14624	(80.9%)
USP	Under-specified referential	"many people"	2446	(10.8%)	1998	(11.0%)
GEN	Generic referential	"extremist groups"	1893	(6.2%)	1407	(7.8%)
NEG	Negatively quantified	"no one"	41	(0.2%)	57	(0.3%)

Table 3.6: *Entity mention classes in the ACE source data.*

Entity mentions are also filtered based on mention class. Table 3.6 contains a list of possible entity mention classes (Label) with a short description (Description), an example (Example) and the number and proportion of occurrences in the ACE 2004 and ACE 2005 data sets. The filtering here removes all entity mentions that are not specific referential, i.e. all mentions that do not refer to a particular, unique object or set of objects in the real world. Next, relation mentions are removed where one of the entity mentions is no longer part of the annotation due to the entity filtering rules. Finally relation mentions in ACE 2004 with relation type DISCOURSE are removed. According to the ACE 2004 Annotation Guidelines for Relation Detection and Characterization (LDC, 2004c): *A* DISCOURSE *relation is one where a semantic part-whole or membership relation is established only for the purposes of the discourse.* Examples include "Many of these people" and "each of whom". In ACE 2004, 279 *discourse* relation mentions were filtered. In ACE 2005, *discourse* relation mentions were discontinued. After filtering, the ACE 2004 data has 13358 entity mentions and 1511 relation mentions (down from 22736 and 4374 respectively in the original source). And the ACE 2005 data has 10345 entity mentions and 975 relation mentions (down from 18086 and 3658).

In the final step in the re-annotation process, entity and relation types are changed to the final schema. This is a simple automatic mapping from the original schema, which serves to simplify the schemas and make them more similar across the development and test sets. Table 3.7 lists the mapping rules, with the first column (#) containing the numeric rule identifier, the second column (Source) containing the types as they are found in the original source data, the third column (Target) containing the types after mapping and the last four columns containing the number and proportion of occurrences in the ACE 2004 and ACE 2005 data sets. In the Source and Target columns, entity and relation type labels prefixed with "T:" are types and labels prefixed with "S:" are sub-types. Rows 1 through 4 of Table 3.7(b), for example, specify that

a) Entity type changes

#	Source	Target	ACE 2004		ACE 2005	
1	T:GPE (Geo-Political)	T:GPL	3262	(87.5%)	3330	(85.3%)
2	T:LOC (Location)		259	(6.9%)	230	(5.9%)
3	T:FAC (Facility)	T:FVW	162	(4.3%)	174	(4.5%)
4	T:VEH (Vehicle)		37	(1.0%)	144	(3.7%)
5	T:WEA (Weapon)		7	(0.2%)	28	(7.2%)

b) Relation type changes

#	Source	Target	ACE 2004		ACE 2005	
1	S:Located,	T:GEN-AFF &	275	(35.1%)	210	(36.3%)
2	S:Near,	S:Located	18	(2.3%)	24	(4.2%)
3	S:Based-In,		106	(13.5%)	NA	NA
4	S:Org-Location[a]		NA	NA	49	(8.5%)
5	S:Cit-Res,[b]	T:GEN-AFF &	70	(8.9%)	NA	NA
6	T:OTHER-AFF,	S:Cit-Res-Rel-Eth	19	(2.4%)	NA	NA
7	T:GPE-AFF &		15	(1.9%)	NA	NA
	S:Other,					
9	S:Cit-Res-Rel-Eth[c]		NA	NA	42	(7.3%)
10	T:ART	T:AGT-ART &	14	(1.8%)	35	(6.1%)
		S:Use-Own-Inv-Mnf[d]				
12	S:Subsidiary	T:PRT-WHL &	80	(10.2%)	81	(14.0%)
		S:Subsidiary				
13	S:Part-Whole,	T:PRT-WHL	187	(23.9%)	NA	NA
14	T:PART-WHOLE		NA	NA	137	(23.7%)

[a]Located, based, headquartered, operates, etc.
[b]Citizen or resident affiliation
[c]Citizen, resident, religious or ethnic affiliation
[d]User, owner, inventor, manufacturer, etc.

Table 3.7: *Changes in ACE entity and relation type schemas. Columns contain the rule identifier (#), the source types (Source), the target types (Target) and counts and percentages for ACE 2004 and ACE 2005. 'T:' and 'S:' indicate relation types and sub-types respectively.*

all relation mentions with sub-type LOCATED, NEAR, BASED-IN or ORG-LOCATION are changed to have type GEN-AFF and sub-type LOCATED. Details of the relation type schema for the original ACE data sets can be found in Appendix B.

Tables 3.8 and 3.9 contain the GRI and GRC type distributions for the final ACE 2004 and ACE 2005 data sets after the full re-annotation process. The first column lists the gold standard type. For GRI, this is a binary distinction between an entity mention pair being in a relation or not being in a relation. For GRC the first column lists the relation type (with super-types typeset in small capital letters). The next seven columns list the entity pair sub-domains. These data subsets are constructed based on four entity types: FACILITY/VEHICLE/WEAPON (FVW or F), GEOGRAPHICAL/POLITICAL/LOCATION (GPL or G), ORGANISATION (ORG or O) and PERSON (PER or P).

The GRI data sets have fewer instances because relation mentions are removed where one or both of the entity mentions are prenominal. This is to make the GRI task consistent with the output from the named entity recognisers used for the extrinsic evaluation in Chapter 6, which does not mark prenominal entity mentions (e.g., "Scottish" in "Scottish National Health Service"). These instances are not filtered for the GRC data in order to maximise the number of data points for evaluation.

3.3.2 Biomedical IE Data: BioInfer

3.3.2.1 Overview

The data for the biomedical domain is derived from the IE corpora that have been prepared and freely distributed as the Bio Information Extraction Resource (BioInfer) corpus by researchers at the University of Turku (Pyysalo et al., 2007).[10] This consists of 1100 sentences that were selected from the PubMed database of biomedical literature[11]. The corpus data was collected by entering known pairs of interacting proteins[12] as PubMed search terms. Resulting abstracts (including titles) were searched for sentences containing mentions of two proteins that are known to interact. The epoch of the resulting corpus includes publication dates up to December 2001, which is when the sentence selection process was carried out.

[10] http://mars.cs.utu.fi/BioInfer/

[11] http://www.ncbi.nlm.nih.gov/pubmed/

[12] Known pairs of interacting proteins were taken from the Database of Interacting Proteins (DIP). See http://dip.doe-mbi.ucla.edu/.

GRI Y/N	F-G	G-G	G-O	G-P	O-O	O-P	P-P
Gold Relation-Forming Pair: Yes	26	159	92	266	42	308	56
Gold Relation-Forming Pair: No	65	1041	749	1805	756	1480	2408
Total	91	1200	841	2071	798	1788	2464

GRC Type	F-G	G-G	G-O	G-P	O-O	O-P	P-P
EMPLOYEE-MEMBERSHIP-SUBSIDIARY							
EMPLOYEE-STAFF				28		275	
EMPLOYEE-EXECUTIVE				88		132	
MEMBER-OF-GROUP					10	70	
OTHER					10	15	
EMPLOY-UNDETERMINED				4		9	
PARTNER					3		
GENERAL-AFFILIATION							
LOCATED	26	9	114	200		3	
CITIZEN-RESIDENT-RELIGION-ETHNIC		6	6	81		5	
PART-WHOLE							
PART-WHOLE		174					
SUBSIDIARY			44	28	28		
PERSONAL-SOCIAL							
BUSINESS							35
FAMILY							15
OTHER							4
AGENT-ARTIFICAT							
USER-OWNER-INVENTOR-MANUFACT	6						
Total	32	189	164	401	51	509	54

Table 3.8: *Relation distributions for GRE news development data (ACE 2004). The first column specifies the relation type and the following columns specify the entity pair sub-domains.*

GRI Y/N	F-G	F-P	G-G	G-O	G-P	O-P	P-P
Gold Relation-Forming Pair: Yes	20	36	87	34	201	119	61
Gold Relation-Forming Pair: No	97	148	1216	658	1405	914	1149
Total	117	59	1303	692	1606	1033	1210

GRC Type	F-G	F-P	G-G	G-O	G-P	O-P	P-P
GENERAL-AFFILIATION							
LOCATED	9	29	9	51	182		
CITIZEN-RESIDENT-RELIGION-ETHNIC					36		3
ORGANISATION-AFFILIATION							
EMPLOYMENT					104	124	
MEMBERSHIP						36	
SPORTS-AFFILIATION						14	
FOUNDER						8	
INVESTOR-SHAREHOLDER						7	
OWNERSHIP						3	
STUDENT-ALUMNUS						3	
PART-WHOLE							
GEOGRAPHICAL	19		100				
SUBSIDIARY				47			
PERSONAL-SOCIAL							
FAMILY							42
BUSINESS							16
LASTING-PERSONAL							10
AGENT-ARTIFACT							
USER-OWNER-INVENTOR-MANUFACT	13	12					
Total	41	41	109	98	322	195	71

Table 3.9: *Relation distributions for GRE news test data (ACE 2005). The first column specifies the relation type and the following columns specify the entity pair sub-domains.*

It should be noted that the targeted selection process means that sentences always have relations, which is not representative of a random sample. However, while the sentences tend to be densely annotated with entity mentions, it is certainly not the case that there is a relation between all pairs of entity mentions. This is illustrated by the figures in Table 3.2 above. While the proportion of the total entity mention pairs in the BioInfer data that are true relation mentions according to the annotation (27.2%)[13] is high compared to ACE 2004 (11.4%) and ACE 2005 (9.1%), it is still quite low, meaning that the relation identification task is still comparatively difficult.

3.3.2.2 Re-Annotation

After converting the BioInfer data to the REXML format (see Appendix A) and performing pre-processing (see the beginning of Section 3.3 above), the annotation is normalised for the GRE task. While the BioInfer data does include some arguably nominal entity mentions (e.g., "complex of birch profilin and skeletal muscle actin), entity mentions are not marked with mention type (let alone full coreference information, entity mention types and entity mention classes as described for the ACE data above) so it is not possible to identify or map nominal entity mentions. As a consequence, the normalisation process for the BioInfer data is simpler than for the ACE data. Nevertheless, it proceeds broadly in the same three steps: 1) mapping entity mentions, 2) filtering relation and entity mentions that are not relevant to the evaluation of the GRE task, and 3) converting entity and relation types to the final schema.

The first step is necessary because of two aspects of the BioInfer annotation that are inconsistent with the GRE task as defined here. First, BioInfer allows n-ary relations over more than two entity mention arguments while the GRE as defined here task only addresses binary relation mentions. Second, the BioInfer annotation sometimes marks part-whole and part-part relation mentions differently depending on their syntactic context. The mapping rules are listed in Table 3.10. The first column (#) lists the rule number. The second column contains a brief rule description on the line where the rule number is given (e.g., **N-ary** $\mapsto$ **Binary**). Below this, the second column contains a list of relation types that are affected in the original BioInfer source data. The third column (Source) contains a count of how many relation mentions in the original data were identified for mapping and the corresponding percentage of total mappable rela-

[13] The total entity mention pairs for this calculation is the total number of pairs of distinct entity mentions that are siblings and occur within the same sentence after the pre-processing described in this section.

#	Description / Relation Type	Source		Target	
1	***N*-ary ↦ Binary**				
	COLOCALIZE	11	(1.9%)	66	(12.9%)
	MUTUALCOMPLEX	9	(1.5%)	39	(7.6%)
	INTERACT	7	(1.2%)	45	(8.8%)
	ATTACH	2	(0.3%)	20	(3.9%)
	BIND	2	(0.3%)	6	(1.2%)
	COEXPRESS	1	(0.2%)	3	(0.6%)
	COPRECIPITATE	1	(0.2%)	3	(0.6%)
	SQSIMILAR	1	(0.2%)	10	(2.0%)
2	**Part-Whole ↦ Part-Part**				
	MEMBER	258	(44.4%)	84	(16.4%)
	CONTAIN	252	(43.4%)	208	(40.7%)
	SUBSTRUCTURE	14	(2.4%)	17	(3.3%)
	F-CONTAIN	13	(2.2%)	6	(1.2%)
	HUMANMADE	10	(1.7%)	4	(0.8%)

Table 3.10: *List of rules for mapping entity mentions in BioInfer. Columns contain the rule identifier (#), the rule description and affected relation types (Description / Relation Type) and the number and percent of relation mentions of the given type in the source (Source) and the mapped (Target) data.*

tion mentions. The fourth column (Target) contains a count of how many new relation mentions are created by the mapping rules and the corresponding percentage of total new relation mentions.

Rule 1 addresses n-ary relation mentions. The solution here is simply to map to binary relation mentions. In the following, for example, the top relation mention with four arguments is replaced by the six distinct binary relation mentions that follow it:

> "Four yeast spliceosomal proteins ([$^{AAC/PTN}$ PRP5], [$^{AAC/PTN}$ PRP9], [$^{AAC/PTN}$ PRP11], and [$^{AAC/PTN}$ PRP21]) interact to promote [$^{AAC/PTN}$ U2 snRNP] binding to [^{NAC} pre-mRNA]."
>
> CHANGE/INTERACT("PRP5", "PRP9", "PRP11", "PRP21")
>
> CHANGE/INTERACT("PRP5", "PRP9")
> CHANGE/INTERACT("PRP5", "PRP11")
> CHANGE/INTERACT("PRP5", "PRP21")
> CHANGE/INTERACT("PRP9", "PRP11")
> CHANGE/INTERACT("PRP9", "PRP21")
> CHANGE/INTERACT("PRP11", "PRP21")

The resulting binary relation mentions may be argued to be incomplete in that they

don't capture the simultaneous interaction between the four proteins.[14] However, they are compatible with the GRE task and the full context of the n-ary relation can still be inferred by looking at the full list of binary relation mentions for a sentence.

Rule 2 addresses relation mentions that are marked differently depending on their syntactic context.[15] Consider the following two sentences:

> "[^{PTN} Smooth muscle talin] prepared from chicken gizzard binds to [^{PTN} skeletal muscle actin]"

> "A binary [^{CPX} complex of [^{PTN} birch [^{PTN} profilin]] and [^{PTN} skeletal muscle actin]] could be isolated by gel chromatography."

The first sentence is annotated with one BIND("Smooth muscle talin", "skeletal muscle actin") relation mention. The second sentence, however, is annotated with two CONTAIN relation mentions where the entity mention "complex of birch profilin and skeletal muscle actin" is the whole and the entity mentions "birch profilin" and "skeletal muscle actin" are the respective parts. In the BioInfer relation type schema, BIND and CONTAIN are defined as follows:

BIND Non-covalent binding (i.e., formation of a complex, association) between the arguments.

CONTAIN A component is part of a complex.

For the annotation to be consistent across the two sentences, the second sentence should also have a relation mention between between "birch profilin" and "skeletal muscle actin". Therefore, a CO-*X* relation mention is added between each entity mention that is annotated as being part of the same whole, e.g. CO-CONTAIN("birch profilin", "skeletal muscle actin").[16]

The next step in the re-annotation process filters certain entity and relation mentions. First, entity mentions are filtered based on top-level entity type. Table 3.11 contains a list of possible types (Label) with a short description (Description), an example (Example) and the number and proportion of occurrences in the BioInfer source data. The examples are with respect to the following sentence:

[14] It is also the case that for some n-ary relation mentions, such as BIND, all pairwise contacts between arguments are not necessarily present. This means that some of the binary relation mentions resulting from the mapping may not be valid. However, the potential noise is considered a reasonable sacrifice in the context of the GRE evaluation here.

[15] Relation mentions sometimes being marked differently depending on their syntactic context may be due to the fact that, as evidenced by early publications (e.g., Pyysalo et al., 2004, 2006), BioInfer was conceived as a corpus for investigating the effects of parsing on the IE task.

[16] As discussed in Footnote 14 above, it may not be strictly true that there are pairwise relations between all entity mentions that are part of the same whole. However, the potential noise is considered a reasonable sacrifice in the context of the GRE evaluation here.

Label	Description	Example	*N*	*%*
PHYSICAL	References to real-world objects	"acanthamoeba profilin", "acanthamoeba actin"	5703	(73.1%)
TEXTBINDING	Minimum text span necessary to resolve entity or relation identity	"inhibits"	1469	(18.8%)
PROCESS	Same as CHANGE sub-tree in relation type schema	"Acanthamoeba actin polymerization"	411	(5.3%)
PROPERTY	Properties associated with entity state (e.g., amount, function)	"rate of Acanthamoeba actin polymerization"	223	(2.9%)

Table 3.11: *Top-level entity types in the full BioInfer entity type schema.*

> "[^{phys} Acanthamoeba [^{phys} profilin]] [^{text} inhibits] the [^{prop} rate of [^{proc} [^{phys} Acanthamoeba [^{phys} actin]] polymerization]] in 50 mM KCl"

Here, all entity mentions that do not have the PHYSICAL super-type according to the entity type schema are removed. This filters entity mentions that do not refer to actual physical objects (i.e., those with entity type PROPERTY, TEXTBINDING or PROCESS). This is compatible with the ontological notion of entity in the context of the GRE task, where an entity mention is assumed to refer to a specific object in the real world.

BioInfer also allows multiple annotations of the same entity mention. This can happen, for example, when a plural pronominal entity mention refers to more than one specific entity mention in the same sentence:

> "[^{gene} 4a] and [^{gene} 4b] are two genes, one of [^{gene} [^{gene} which]] codes for the proposed [^{ptn} phosphoprotein] [^{ptn} P]"

where "which" refers back to "4a" and to "4b". Here, mentions that do not take part in a relation are removed until there is only one left. As the last step of filtering based on entity types, relation mentions are removed where one of the entity mentions is no longer part of the annotation due to the entity filtering rules.

Relation mentions are also filtered based on type. First, relation mentions with type REL-ENT are removed. These are BioInfer relations where an unnamed entity mention refers to a named entity mention, e.g.:

> "PRP incubated with [^{ptn} IL-6] showed a [^{amount} dose] dependent increase in [^{protein} TXB2]"

where the REL-ENT("dose", "IL-6") relation mention indicates that dose refers to dose of IL-6. The original BioInfer data contains 50 REL-ENT relation mentions. After

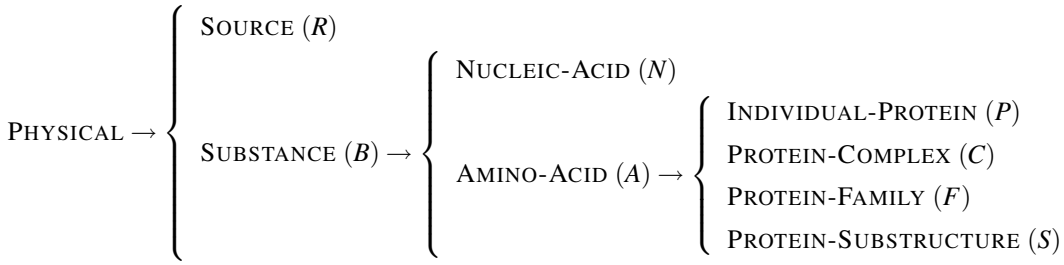


Figure 3.6: *Simplified entity type schema for BioInfer.*

filtering, the BioInfer data has 5800 entity mentions and 2116 relation mentions (down from 7818 and 3020 respectively in the original source).

In the final step of the re-annotation process, entity and relation types are changed to the final schema. This is a matter of choosing a level in the full relation type schema from the source BioInfer data that gives several entity pair sub-domains with a sufficient number of relation types and instances for evaluation of the GRC task. Figure 3.6 contains a simplified version of the entity type schema (see Pyysalo et al. (2007) for details of full entity type schema). The entity pair subset for each relation mention is determined by choosing the lowest level in this schema where the types of the entity mentions are siblings. For example, the sub-domain for a relation-forming pair consisting of an INDIVIDUAL-PROTEIN (P) entity mention and a PROTEIN-COMPLEX (C) entity mention would be P-C. For a pair consisting of a SOURCE entity mention and an INDIVIDUAL-PROTEIN entity mention – with parent type SUBSTANCE (B), however, the sub-domain would be R-B. The relation type for the GRC task is simply the second-level type from the full relation schema (see Appendix B), i.e. one of CAUSAL, PART-OF, OBSERVATION or IS-A.

Table 3.12 contains the GRI and GRE type distributions for the final BioInfer data set after the full re-annotation process.[17] The first column lists the gold standard type. For GRI, this is a binary distinction between an entity mention pair being in a relation or not being in a relation. For GRC, the first column lists the relation type (with super-types typeset in small capital letters). The next seven columns list the entity pair sub-domains. These data subsets are constructed as described above based on the eight entity types under PHYSICAL in Figure 3.6.

The GRC data subsets have fewer instances because a number of relation mentions

[17] Including the removal of relation mentions that are not between distinct siblings (discussed at the beginning of Section 3.3).

GRI Y/N	A-N	P-P	P-C	P-F	P-S	N-N	R-B
Gold Relation-Forming Pair: Yes	43	942	130	193	130	49	104
Gold Relation-Forming Pair: No	182	2450	183	521	229	362	325
Total	225	3392	313	714	359	411	429

TYPE	A-N	P-P	P-C	P-F	P-S	N-N	R-B
CAUSAL	12	469	27	13	100	9	69
PART-OF	3	43	103	174	12	10	4
OBSERVATION		134					16
IS-A	27	48			14	14	
Total	42	694	130	187	126	33	89

Table 3.12: *Relation distributions for GRE biomedical test data (BioInfer). The first column specifies the relation type and the following columns specify the entity pair sub-domains.*

that have vague or undetermined types are ignored for the relation characterisation experiments (but not the relation identification experiments). These include the following (with their BioInfer definitions):

CORELATE A general, unspecified co-relation between the arguments.

HUMANMADE A relationship that is forced or caused by human intervention. The actual type of the relationship is not stated but is one of the types in the schema.

RELATE A general, unspecified, non-directional relationship used when no details of the relationship are known.

CO-* The relations created by Rule 2 for mapping entity mentions (see Table 3.10 above).

3.4 Intrinsic Evaluation

In the current work, the intrinsic evaluation focuses on the GRI task (Chapter 4) and on the GRC clustering sub-task (Chapter 5). The evaluation of end-to-end GRE is addressed in the extrinsic evaluation (see Chapter 6). Existing evaluation scripts for supervised RE (e.g., the scripts written for the ACE data) are not applicable here as they assume an output where relations are labelled according to a predetermined relation type schema. The output of GRE is simply a partition over the data (clustering),

optionally including cluster labels. In the remainder of this section, an intrinsic evaluation for GRE is described. First, Section 3.4.1 defines some standard evaluation measures. Next, Sections 3.4.2 and 3.4.2 respectively describe the evaluations for GRI and GRC.

3.4.1 Standard Evaluation Measures

A number of the approaches that are discussed in the following sections rely on some commonly used evaluation formulae from the IR and NLP literature (e.g., Manning and Schütze, 1999; Tan et al., 2005; Manning et al., 2008), namely precision, recall and f-score. Precision and recall are commonly defined in terms of a contingency table like the following:

	Gold	
System	Yes	No
Yes	*tp*	*fp*
No	*fn*	*tn*

where *tp* consists of true positives (instances correctly classified as belonging to the target class), *fp* consists of false positives (instances incorrectly classified as belonging to the target class), *fn* consists of false negatives (instances incorrectly classified as not belonging to the target class), and *tn* consists of true negatives (instances correctly classified as not belonging to the target class).

Given the values from the contingency table above, precision P and recall R for given class are defined as:

$$Precision = P = \frac{tp}{tp+fp} \qquad Recall = R = \frac{tp}{tp+fn} \tag{3.1}$$

P is the proportion of correct instances among all of the instances that the system assigned to the target class. R is the proportion of correct instances among all of the instances that the system should have assigned to the target class.

F-score is a measure that combines precision and recall using the harmonic mean (Manning et al., 2008):

$$F(P,R,\beta) = \frac{(\beta^2+1)PR}{\beta^2 P+R} \tag{3.2}$$

where β is a factor that determines the relative weighting of precision and recall. Convention is to equally weight P and R by using $\beta = 1$, which allows the definition to be simplified to:

$$F(P,R) = \frac{2PR}{P+R} \tag{3.3}$$

System Ent Pairs		
ID	Entity 1	Entity 2
s1	“Toefting”	“Bolton”
s2	“Toefting”	“Hamburg”
s3	“Hamburg”	“Bolton”

Gold Ent Pairs		
ID	Entity 1	Entity 2
g1	“Toefting”	“Bolton”
g2	“Toefting”	“Hamburg”

Table 3.13: *Example input to GRI evaluation.*

This is sometimes referred to as the balanced f-score. F-score measures have the advantage of being widely used and well understood within the NLP community. Furthermore, they are useful for analysis as it makes it possible to separate and study the interaction between Type I (precision) and Type II (recall) errors.

3.4.2 Generic Relation Identification

3.4.2.1 GRI Evaluation: Input and Output

The output of the GRI system is a list of entity mention pairs that are considered to form a relation. This is the primary input to the GRI evaluation. The secondary input is the list of true relation-forming entity mention pairs from the gold standard data. Take the following sentence:

> “[^{per} Toefting] transferred to [^{org} Bolton] from [^{org} Hamburg].”

As illustrated in Table 3.13, a system might predict three relation mentions while the gold standard has only two relation mentions. This example will be used to illustrate the following explanation of GRI evaluation.

3.4.2.2 GRI Evaluation: Precision and Recall

In the context of the GRI task, tp_{gri} is the number of entity mention pairs identified by the system that are true according to the gold standard annotation. Precision is defined as:

$$Precision_{gri} = P_{gri} = \frac{tp_{gri}}{tp_{gri} + fp_{gri}} \tag{3.4}$$

where the denominator $(tp_{gri} + fp_{gri})$ is the total number of entity mention pairs identified by the system. Recall is defined as:

$$Recall_{gri} = R_{gri} = \frac{tp_{gri}}{tp_{gri} + fn_{gri}} \tag{3.5}$$

where the denominator ($tp_{gri} + fn_{gri}$) is the total number of entity mention pairs according to the gold standard annotation. Taking the example GRI evaluation input in Table 3.13, tp_{gri} is equal to two as the system entity mention pairs s1 and s2 are the same as gold standard entity mention pairs g1 and g2 respectively. The total number of system entity mention pairs is three so $P_{gri} = 2/3 = 0.667$ and the total number of gold standard entity mention pairs is two so $R_{gri} = 2/2 = 1.000$. Combined precision and recall score is computed using the balanced f-score (Equation 3.3) as $F = \frac{2*0.667*1}{0.667+1} = 0.800$.

3.4.2.3 GRI Evaluation: Discussion

The simple GRI evaluation scheme described is an improvement over previous evaluations of GRI, in that it defines a combined measure of precision and recall with respect to an established gold standard. This evaluation is used here to develop and evaluate approaches to the automatic identification of relation-forming entity mention pairs with respect to an established RE gold standard. This evaluation does not address ranking of entity pairs, which is not necessary for the GRC task in isolation. Ranking approaches are considered in the analysis for the experiments in Chapter 4 and in the extrinsic evaluation in Chapter 6.

3.4.3 Generic Relation Characterisation

3.4.3.1 GRC Evaluation: Input and Output

The output of a clustering system for GRC is an automatically induced partition that groups entity mention pairs with respect to relation type.[18] One option would be to use internal measures of cluster quality from the clustering algorithm (e.g., the I_2 criterion function defined in Section 5.2.1.3). However, while these may be useful for comparisons among similar systems such as for optimisation, they do not validate the system with respect to any external objective notion of what is correct. Therefore, they are not necessarily a reliable measure for comparison of more heterogeneous systems and are therefore not generally useful for reporting.

The evaluation here uses external measures of clustering accuracy which compare system output to a gold standard. Thus, the primary input to the clustering evaluation

[18] While clustering output can also be a dendrogram in the case of hierarchical clustering, it is straightforward to use the dendrogram to define a flat partition by cutting the tree at the point that gives the appropriate number of clusters. Section 5.2.1.3 describes this distinction in more detail and motivates the use of hierarchical clustering.

R	S	Entity 1	Entity 2	System Cluster	Gold Relation Type
1	1	"martha stewart"	"m.s. living omnimedia"	c1	EMPLOY-EXECUTIVE
2	2	"Toefting"	"Bolton"	c2	SPORTS-AFFILIATION
3	2	"Toefting"	"Hamburg"	c2	SPORTS-AFFILIATION
5	3	"David Murray"	"Amidu Berry"	c1	BUSINESS

Table 3.14: *Example input to GRC evaluation.*

algorithm is the partition defined by the system output. And the secondary input is a gold standard partition of the same data as defined by the relation type annotation. Take the following three sentences for example:

1 "[^{per} martha stewart]'s company is registered as [^{org} m.s. living omnimedia]"
2 "[^{per} Toefting] transferred to [^{org} Bolton] from [^{org} Hamburg]."
3 "[^{per} David Murray] recruited [^{per} Amidu Berry]."

These sentences contain four PERSON-ORGANISATION (*per-org*) entity mention pairs listed in Table 3.14. The first column (R) contains the relation identifier. The second column (S) contains the sentence identifier, which links the entity mention pairs back to the source sentences above. The third (Entity 1) and fourth (Entity 2) columns contain the entity mentions. The fifth column (System Cluster) contains the cluster identifier from the system output. And the sixth column (Gold Relation Type) contains the true relation type from the gold standard annotation.

In this example, the system posits two clusters: c1 (of which Relation Mentions 1 and 5 are instances) and c2 (of which Relation Mentions 2 and 3 are instances). According to the gold standard annotation, however, there are three relation types: EMPLOY-EXECUTIVE (of which Relation Mention 1 is the sole instance), SPORTS-AFFILIATION (of which Relation Mentions 2 and 3 are instances) and BUSINESS (of which Relation Mention 5 is an instance).

3.4.3.2 GRC Evaluation: Precision and Recall

In the context of the clustering task, an intuitive definition of precision and recall is in terms of sets. Standard notation used for definition of GRC evaluation measures is summarised in Table 3.15. For the purposes of these definitions, C and L represent partitions over the data D. C represents a partition defined by the output of the clustering system and L represents a partition defined by the gold standard labelling. Indices on

D	Data points (i.e. clustering instances)
C	Partition defined by clustering system output
C_i	Output cluster indexed by i
$\|C_i\|$	Number of data points in cluster C_i
L	Partition defined by gold standard labelling
L_j	Gold standard class indexed by j
$\|L_j\|$	Number of data points in class L_j

Table 3.15: *Standard notation for GRC evaluation measures.*

the partition variables, e.g. C_i and L_j, indicate subsets of data points corresponding to individual clusters (or classes).

In these terms, precision measures the proportion of instances in system cluster C_i that are correct with respect to gold standard class L_j:

$$Precision(C_i, L_j) = P(C_i, L_j) = \frac{|C_i \cap L_j|}{|C_i|} \tag{3.6}$$

Recall measures the proportion of instances in gold standard class L_j that the system correctly grouped together in cluster C_i:

$$Recall(C_i, L_j) = R(C_i, L_j) = \frac{|C_i \cap L_j|}{|L_j|} \tag{3.7}$$

These equations require a mapping between system clusters and gold standard classes. Possible mappings will be described in the following sections.

3.4.3.3 GRC Evaluation: Purity And Inverse Purity

Purity is a standard measure of cluster performance against a gold standard that is analogous to precision and can be computed in reverse to obtain a measure analogous to recall. Purity is calculated at the cluster level as:

$$Purity(X, Y, i) = \max_j Precision(X_i, Y_j) \tag{3.8}$$

where X and Y define partitions over the data D. Overall purity (i.e. the purity of the overall clustering) is defined as the weighted mean of the cluster-level purity scores:

$$Purity(X, Y) = \sum_i \frac{|X_i|}{|D|} Purity(X, Y, i) \tag{3.9}$$

Standard purity $Purity(C,L)$ measures the extent to which the clusters contain objects of a single class (Tan et al., 2005).

Inverse purity $Purity(L,C)$ is sometimes reported in conjunction with purity, e.g. Artiles et al. (2007). This can then be combined using the harmonic mean to give a balanced f-score for the overall clustering:

$$F_{pur}(C,L) = F(Purity(C,L), Purity(L,C)) \tag{3.10}$$

It is not possible to calculate an f-score at the cluster level as there is no explicit mapping between clusters and classes.

Hasegawa et al. (2004) and Zhang et al. (2005) use a variation on purity-based accuracy ($F_{n:1}$) to evaluate the relation discovery clustering task, which is defined in terms of a many-to-one mapping from clusters to gold standard classes that is calculated using the cluster-level purity measure in Equation 3.8. To define the measure using the precision and recall formulae from Section 3.4.1, the mapping can be formulated as:

$$\Omega(i) = \arg\max_j |C_i \cap L_j| \tag{3.11}$$

which maps cluster C_i to the gold standard class $L_{\Omega(i)}$ with the highest overlap. The advantages of this measure are twofold. First, it possible to compute cluster-level accuracy scores. Second, the mapping is simple and efficient to compute.

3.4.3.4 GRC Evaluation: Chen et al. (2005)

Chen et al. (2005) and Chen et al. (2006) use precision and recall measures for relation discovery based on an optimal mapping $\hat{\Omega}$ from gold standard classes to clusters:

$$\hat{\Omega}(C,L) = \arg\max_{\Omega} \sum_i \phi(C,L,i,\Omega) \tag{3.12}$$

where $\phi(C,L,i,\Omega) = |L_i \cap C_{\Omega(i)}|$. Chen et al. constrain their mapping to be one-to-one. If there are more classes than clusters, then some classes are left unaligned (likewise if there are more clusters). Therefore the measure penalises systems that propose too many or too few clusters.

The one-to-one mapping constraint also allows the f-score to be calculated for individual cluster/class pairs. One-to-one precision $P_{1:1}$ and recall $R_{1:1}$ are defined at the cluster level as:

$$P_{1:1}(C,L,i,\Omega) = Precision(C_{\Omega(i)}, L_i) \tag{3.13}$$

$$R_{1:1}(C,L,i,\Omega) = Recall(C_{\Omega(i)}, L_i) \tag{3.14}$$

The cluster-level balanced f-score is computed as:

$$F_{1:1}(C,L,i,\Omega) = F(P_{1:1},R_{1:1}) \tag{3.15}$$

where the C, L, i and Ω parameters to precision and recall are omitted for readability. Overall precision and recall can computed like overall purity,[19] from which an overall f-score is computed in the standard way.

Like the $F_{n:1}$ measure in the previous section, the $F_{1:1}$ measure here has the advantage of providing cluster-level accuracy scores. In contrast to $F_{n:1}$ measure, the $F_{1:1}$ measure provides a stricter accuracy measure where classes cannot be aligned with multiple clusters. The disadvantage of the $F_{1:1}$ measure is that it can be prohibitively expensive to perform an exhaustive search of one-to-one mappings. However, a simple greedy search through possible alignments with a beam of width five has linear time and space complexity and provides a reasonable approximation.[20]

Purity-based accuracy measures based on one-to-one mappings are well attributed in the NLP and general clustering literatures, e.g. for evaluation of unsupervised part-of-speech tagging (Haghighi and Klein, 2006; Johnson, 2007) and coreference resolution (Popescu-Belis and Robba, 1998; Trouilleux et al., 2000; Luo, 2005).

3.4.3.5 GRC Evaluation: Pairwise Precision and Recall

While the previous measures require an explicit mapping between the clustering output and the gold standard labelling, there is another group of measures that does not. These are collectively referred to as pairwise measures as they are based on the distribution of pairs of data points and are computed by calculating the agreement between pairs of data points, i.e. whether they are grouped together in both the system clustering and the gold standard. These are often defined in terms of the following contingency table:

	Classes	
Clusters	Same	Diff.
Same	a	b
Diff.	c	d

where a corresponds to the number of pairs of data points in the same cluster and in the same class, b corresponds to the number of pairs in the same cluster but in different

[19] Chen do not specify how they compute their overall measures. My implementation weights both precision and recall by cluster size.

[20] Exact solutions are also possible, e.g. Luo (2005) describes an approach that uses the Kuhn-Munkres algorithm and has polynomial time complexity.

classes, c corresponds to the number of pairs in different clusters but in the same class, and d corresponds to the number of pairs in different clusters and different classes.

Given the values of the contingency table above, pairwise precision P_{pw} and recall R_{pw} for the overall clustering are calculated as:

$$P_{pw}(a,b) = \frac{a}{a+b} \qquad R_{pw}(a,c) = \frac{a}{a+c} \tag{3.16}$$

A balanced f-score for the overall clustering is calculated as:[21]

$$F_{pw} = F\left(P_{pw}, R_{pw}\right) \tag{3.17}$$

where the a, b and c parameters are omitted for readability. This is very similar to other pairwise index measures widely used in the clustering literature such as Rand, Jaccard and Fowlkes-Mallows.[22] Rand Index $(a+d)/(a+b+c+d)$ is the same as accuracy (sometimes used in the NLP literature (e.g., Manning and Schütze, 1999, p269)). It is considered a bad measure because fn counts inflate scores and make it difficult to distinguish between systems. The Jaccard Coefficient $a/(a+b+c)$ and Fowlkes-Mallows Index $a/\sqrt{(a+b)(a+c)}$ address this problem. However, they are less familiar to the NLP research community and do not separate Type I and Type II errors like F_{pw}.

A disadvantage of pairwise f-score (and other pairwise measures) is that it is not possible to compute cluster-level scores. Nevertheless, pairwise f-score is commonly used for document clustering tasks (e.g., Basu et al., 2004; Liu et al., 2007a). It has also been used to evaluate automatic lexical acquisition tasks such as grouping adjectives by meaning (Hatzivassiloglou and McKeown, 1993) and induction of verb frames and classes (Schulte im Walde, 2003).

3.4.3.6 GRC Evaluation: Discussion

The $n:1$ and $1:1$ accuracy measures discussed here (Sections 3.4.3.3 and 3.4.3.4 respectively) provide intuitive and efficient performance measures based on mappings to gold standard relation type annotation. This makes it possible to perform rapid prototyping and evaluation with respect to an explicit mapping between system clusters and

[21] An alternative combined pairwise score is normalised mutual information (e.g., Strehl and Ghosh, 2003; Manning et al., 2008). Preference is given here to pairwise f-score based on its interpretability and its familiarity in NLP. Exclusive use of either pairwise f-score or normalised mutual information is also supported empirically by an evaluation of document clustering that shows identical results using both measures (Basu et al., 2004).

[22] See e.g. Halkidi et al. (2001); Knowles and Kell (2005); Tan et al. (2005) for an overview of related pairwise index measures for clustering evaluation.

gold standard classes. The pairwise accuracy measure (Section 3.4.3.5), on the other hand, does not require an explicit mapping between clusters and gold standard classes, which eliminates the alignment procedure and associated parameters from the evaluation algorithms. In addition, the pairwise measures are based on pairs of clustering instances, which is a natural level for error analysis of clustering. The disadvantage of the pairwise accuracy measures is that they do not define cluster-level scores.

In the current work, the one-to-one f-score ($F_{1:1}$) and the pairwise f-score (F_{pw}) are used together for development and evaluation. The $F_{1:1}$ measure is used in Chen et al.'s closely related work on relation discovery and it is also well attributed in the literature for related NLP tasks like coreference resolution and unsupervised part-of-speech tagging. The F_{pw} is also well attributed in the clustering literature, but has not previously been used for evaluation of the GRC task. Furthermore, instance pairs are used for error analysis of the clustering output and it is therefore useful to have a related evaluation measure for consistency.

3.4.4 Statistical Significance Testing

Most of the experiments here use paired Wilcoxon signed ranks tests (e.g., Coolican, 2004) across entity pair sub-domains to check for significant differences between systems. This is non-parametric analogue of the paired t test. The t test is not used here because it assumes that the underlying distribution is normal, which is not the case for all of experimental results here. 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.

3.5 Summary

This chapter addressed shortcomings in the literature with respect to standardised task definitions and evaluation. In particular, previous approaches have adopted different task definitions and evaluation approaches making meaningful comparison across approaches difficult. First, in Section 3.2, a combined framework for generic identification and characterisation was presented. Second, in Section 3.3, two standard and publicly available IE corpora were described along with a three-stage process (re-factoring, pre-processing, re-annotation) for adapting these corpora to the GRE task. From these corpora, three comparable data sets are derived: 1) the ACE 2004 data is used for de-

velopment in the news domain; 2) the ACE 2005 data is used for testing in the news domain; and 3) the BioInfer data is used for testing in the biomedical domain. This allows evaluation across distinct epochs within the news domain and evaluation of the claim of modification-free domain domain adaptation.

Finally, in Section 3.4, detailed frameworks were presented for evaluating generic relation identification characterisation with respect to gold standard relation extraction data. Unlike many of the previous evaluations, these frameworks are fully automatic and do not require human judgements of system output. Furthermore, the multiple entity pair sub-domains provide a natural level for statistical significance testing which is completely absent from previous work on the GRI and GRC tasks. The result is a rigorous experimental design with held-out evaluation data sets in multiple domains and the use of paired Wilcoxon signed ranks tests to quantify significant differences across entity pair sub-domains.

Chapter 4

Generic Relation Identification

Experiments are reported that address the generic relation identification task, comparing window-based models (e.g., setting a threshold on the number of intervening tokens) for establishing entity mention pair co-occurrence. In related work, co-occurrence windows have been defined in terms of sentence boundaries or intervening token counts. Here, a new approach is introduced that defines windows based on syntactic governor-dependency paths. Experimental results suggest that a combined model based on intervening words and dependency paths is preferable as it is better in terms of recall while being statistically indistinguishable in terms of precision and f-score. Furthermore, the accuracy of optimised models is shown to be comparable across domains. Importantly for applications of GRE, analysis demonstrates that many false positive relations are actually implicit relations that are not part of the gold standard relation schemas. Analysis also suggests that the f-score of high-recall models could be improved using false positive filters.

4.1 Introduction

Related approaches to generic relation identification (GRI) have previously been spread across the distinct literatures of entity association mining and relation discovery (discussed in Chapter 2). This chapter contains a specification of the task in terms of model parameters that incorporate aspects of these various approaches. Figure 4.1 contains an overview of the GRI task, which is split into two main sub-tasks. The input is a collection of natural language documents with entity mentions identified.[1] The first sub-task has the goal of identifying relation-forming entity mention pairs and outputs a

[1]For the evaluation here, the input includes gold standard entity annotation as discussed in Chapter 3. The ACE data input only includes entity mentions that are named or pronominal as discussed at the end of Section 3.3.1.2.

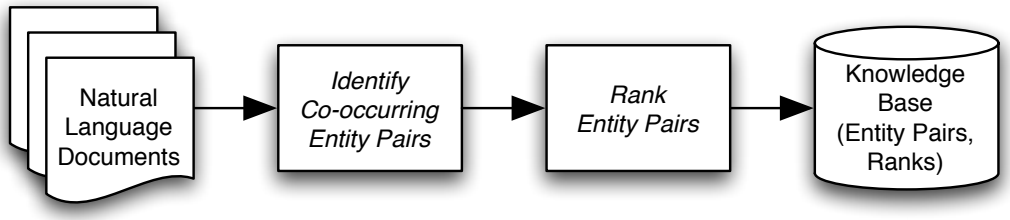


Figure 4.1: *Overview of GRI sub-tasks.*

list of co-occurring entity mention pairs. The second sub-task has the goal of applying a ranking over co-occurring pairs that indicates e.g. the strength of association. The primary experimental focus of this chapter is on the identification sub-task, which can be evaluated with respect to gold standard data. The analysis in this chapter (Section 4.5.4) and the extrinsic evaluation (Chapter 6) address the ranking sub-task.

As discussed in Chapter 2, previous GRI work has largely failed to use standardised data or evaluation measures and has provided little comparison across approaches. This chapter employs a principled framework for evaluation (introduced in Chapter 3) that makes use of gold standard relation extraction data to optimise and evaluate GRI models. News data from the ACE shared tasks is used for development and for testing on a held-out evaluation set in the same domain. The presence of double annotation in the ACE 2005 data makes it possible to compute a human upper bound for the GRI task. Biomedical data from BioInfer is also used, which allows assessment of model consistency across application domains.

Previous GRI work has relied extensively on co-occurring entity mention windows defined in terms of intervening token thresholds or in terms of natural boundaries like documents or sentences. Some constraints on intervening tokens have also been suggested. Filatova and Hatzivassiloglou (2003), for example, require a verbal connector in the intervening context. And, Zhang et al. (2005) require a parse that spans both entity mentions and thus includes a path connecting them. In the literature on supervised (including rule-based) relation extraction (RE), features based on parse trees have been used successfully to learn extractors for specific tasks and domains (e.g., Zelenko et al., 2002; Bunescu et al., 2004; Daraselia et al., 2004; Harabagiu et al., 2005; Riedel and Klein, 2005; Zhou et al., 2005; Fundel et al., 2007; Liu et al., 2007b). However, beyond requiring a spanning parse tree, 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.

A detailed description of previous work can be found in Chapter 2. This chapter begins with a description of the setup for experimental evaluation in Section 4.2. Next, Section 4.3 contains a specification of the models that are compared here. Sections 4.4.1 through 4.4.4 contain experimental results and discussion. Finally, Section 4.5 contains a detailed analysis of the experimental results.

4.2 The Task: Experimental Setup

4.2.1 GRI Based on Co-occurrence Windows

Previous work on GRI has been based on defining a window and counting all entity mention pairs within that window as co-occurring or relation-forming. These approaches can be generalised in terms of the GENERICRELATIONID algorithm in Figure 4.2. 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 4.2 includes the $\text{ISPAIR}_{baseline}$ function as an example, which simply counts all pairs of entity mentions occurring in the same sentence as relation-forming pairs. 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 from E, which is assumed to be sorted in terms of the order of occurrence (i.e., increasing entity mention start location as the primary sort index and decreasing entity mention end location as the secondary sort index, which places embedded entity mentions after their parent embedding entity mentions). 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. These are defined in Section 4.3 below.

4.2.2 Data and Evaluation

The evaluation uses news data from the Automatic Content Extraction (ACE) 2004 and 2005 shared tasks and biomedical data derived from the BioInfer corpus (see Chapter 3 for details of data sets and preparation). 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. The evaluation measure used here is the balanced

```
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        i ← i + 1
9  return P
```

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

Figure 4.2: *Algorithm for generic relation identification with baseline function for identifying co-occurring entity mention pairs.*

f-score described in Chapter 3 (Section 3.4.2). This is calculated with respect to the gold standard data where precision is defined as the number of correct entity mention pairs divided by the number of predicted entity mention pairs and recall is defined as the number of correct entity mention pairs divided by the number of gold standard entity mention pairs.

An important aspect of the evaluation here is the introduction of an upper bound based on human agreement. The ACE 2005 data includes markup from two human annotators and a final adjudicated version of the markup, which makes it possible to compute inter-annotator 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 optimised one-to-one mapping based on maximal character overlap between entity mention strings LDC (2004a). Given this mapping, it is possible to determine for each putative entity mention pair whether the annotators marked a relation mention. Figure 4.3 contains the ISPAIR$_{human}$ function which returns true if a relation mention between entity mentions i and j is marked by the given annotator. This assumes that annotated relation mentions can be read from a two-dimensional matrix A_l that contains entries for all relation-forming entity mention pairs in the markup from annotator l. This matrix is read from a file containing relation mention markup over the mapped entity mention identifiers.

Table 4.1 contains precision (P), recall (R) and f-score (F) results for the individual

```
ISPAIR_human : i, j, A_l
1  if exists(A_l[i, j])
2        return true
3  else
4        return false
```

Figure 4.3: *Function for computing relation identification values for annotators.*

	P	R	F
Human 1	0.888	0.697	0.780
Human 2	0.924	0.653	0.761
Mean	0.906	0.675	0.773

Table 4.1: *Precision (P), recall (R) and f-score (F) results for human annotators against adjudicated gold standard.*

human annotators when compared to the final adjudicated data set. The first two rows contain the individual annotator results and the bottom row contains the mean of the two individual annotators. Interestingly, the annotators have high agreement with the adjudicated data set in terms of precision and lower agreement in terms of recall. 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. This is a good score with respect to other relation extraction annotation efforts that report inter-annotator agreement (e.g., Alex et al. (2008b) report f-score agreement of 0.761 for a protein-protein interaction corpus and 0.741 for a tissue expression corpus).

4.3 Models

In this section, the different models of entity mention co-occurrence used to extract relation-forming pairs are described in detail. Figure 4.4 contains an example sentence and the entity mention pairs extracted by various possible systems based on the co-occurrence models used here. The first row contains the example sentence where entity mention starts and ends are marked with square brackets and the entity mention type is

Example Sentence	[^{place} American] saxophonist [^{person} David Murray] recruited [^{person} Amidu Berry] and DJ [^{person} Awadi] from [^{organisation} PBS].
Baseline	{<American,David_Murray>, <American,Amidu_Berry>, <American,Awadi>, <American,PBS>, <David_Murray,Amidu_Berry>, <David_Murray,Awadi>, <David_Murray,PBS>, <Amidu_Berry,Awadi>, <Amidu_Berry,PBS>, <Awadi,PBS>}
Event	{<American,Amidu_Berry>, <American,Awadi>, <American,PBS>, <David_Murray,Amidu_Berry>, <David_Murray,Awadi>, <David_Murray,PBS>}
Toks (t=0)	{}
Toks (t=2)	{<American,David_Murray>, <David_Murray,Amidu_Berry>, <Amidu_Berry,Awadi>, <Awadi,PBS>}
Toks (t=5)	{<American,David_Murray>, <American,Amidu_Berry>, <David_Murray,Amidu_Berry>, <David_Murray,Awadi>, <Amidu_Berry,Awadi>, <Amidu_Berry,PBS>, <Awadi,PBS>}
Deps (d=0)	{<American,David_Murray>, <Amidu_Berry,Awadi> <Amidu_Berry,PBS>, <Awadi,PBS>}
Deps (d=1)	{<American,David_Murray>, <David_Murray,Amidu_Berry>, <David_Murray,Awadi>, <Amidu_Berry,Awadi>, <Amidu_Berry,PBS>, <Awadi,PBS>}
Comb (t=2,d=0)	{<American,David_Murray>, <David_Murray,Amidu_Berry>, <Amidu_Berry,Awadi>, <Amidu_Berry,PBS>, <Awadi,PBS>}
Gold Standard	{<American,David_Murray>, <David_Murray,Amidu_Berry>, <David_Murray,Awadi>, <Amidu_Berry,PBS>, <Awadi,PBS>}

Figure 4.4: *Example sentence and extracted entity mention pairs corresponding to various co-occurrence models: Baseline, atomic events (Event), intervening tokens (Toks), dependency paths (Deps) and Gold Standard. Where relevant, window size is specified in parentheses under the model type.*

indicated by the superscript text to the right of the opening bracket. In the remaining rows of the table, the model type is specified in the first column and the set of entity mention pairs extracted by that model is given in the second column. The models are described in detail in the following sections.

4.3.1 Baseline

The baseline system uses the IsPair$_{baseline}$ function defined in Figure 4.2. As mentioned, this counts all pairs of entity mentions occurring in the same sentence as relation-forming pairs. This is the same co-occurrence model used by Smith (2002). An example sentence and the relation-forming pairs extracted by the baseline model can be seen in Figure 4.4. The baseline model has perfect recall with respect to the gold standard relation mention set in the last column but it also generates the most precision errors of all the models. For example it posits a relation mention between "American" and "Amidu Berry", which is clearly not supported by the semantics of the sentence and is not actually true in the world as Amidu Berry is from Senegal.

4.3.2 Atomic Events

The results here are also compared to a system based on the approach to identifying atomic events from Filatova and Hatzivassiloglou (2003). This uses the IsPair$_{event}$ function defined in Figure 4.5. This 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. The IsPair$_{event}$ function assumes access to a function (*intervening-connectors*) that returns the set of connectors that occur between the two entity mentions indexed by i and j. An example sentence and the relation-forming pairs extracted by the event-based model (Event) can be seen in Figure 4.4. While this example suggests that the event-based model should not be expected to have high recall for relation identification, the model is useful for comparison. Furthermore, despite having low recall, it might provide a method for identifying long-distance verbal relation mentions (explored in Section 4.4.3 below).

4.3.3 Intervening Token Windows

The next model of entity mention co-occurrence is based on intervening token windows (Toks). It uses the IsPair$_{toks}$ function defined in Figure 4.6. This counts all pairs of

$\text{IsPair}_{event} : i, j$

1 **if** $sent(i) = sent(j)$ **and** $count(intervening\text{-}connectors(i, j)) \geq 1$
2 **return** $true$
3 **else**
4 **return** $false$

Figure 4.5: *Function for GRI based on Filatova and Hatzivassiloglou (2003) atomic events.*

entity mentions that 1) occur in the same sentence and 2) have t or fewer intervening tokens. This assumes access to a function (*intervening-tokens*) that returns the set of tokens that occur between the two entity mentions indexed by i and j. For the current work, stop word tokens and entity mention word tokens (from other entity mentions than the two under consideration) are included in the count of intervening tokens. Most previous GRI work has used some variant of this model. Hasegawa et al. (2004), for example, use the IsPair_{toks} function with the intervening token threshold t set to 5. However, Hasegawa et al. do not explicitly motivate this choice.

An example sentence and the relation-forming pairs extracted by the intervening token (Toks) model with various settings of the threshold t can be seen in Figure 4.4. The second, third and fourth rows correspond to models with t set to 0, 2 and 5 respectively. The t=0 system is the worst in terms of recall on the example sentence as there are no entity mention pairs with zero intervening tokens. The t=2 system does well in terms of precision, generating one false positive relation mention between "Amidu Berry" and "Awadi".[2] It does less well in terms of recall due to missing the <David_Murray, Awadi> and <Amidu_Berry, PBS> relation mentions. The t=5 system achieves perfect recall on the example sentence, but generates two false positive relation mentions (i.e., <American, Amidu_Berry> and <Amidu_Berry, Awadi>).

Figure 4.7 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 (i.e., $p \leq 0.05$) are underlined. The results

[2]It could be argued that there is an implicit relation mention in the example sentence from Figure 4.4 between "Amidu Berry" and "Awadi" because the are both members of "PBS". However, implicit relation mentions are not annotated in the ACE corpora.

IsPair$_{toks}$: i, j, t

1 **if** $sent(i) = sent(j)$ **and** $count(intervening\text{-}tokens(i, j)) \leq t$

2 **return** $true$

3 **else**

4 **return** $false$

Figure 4.6: *Function for GRI based on intervening token windows.*

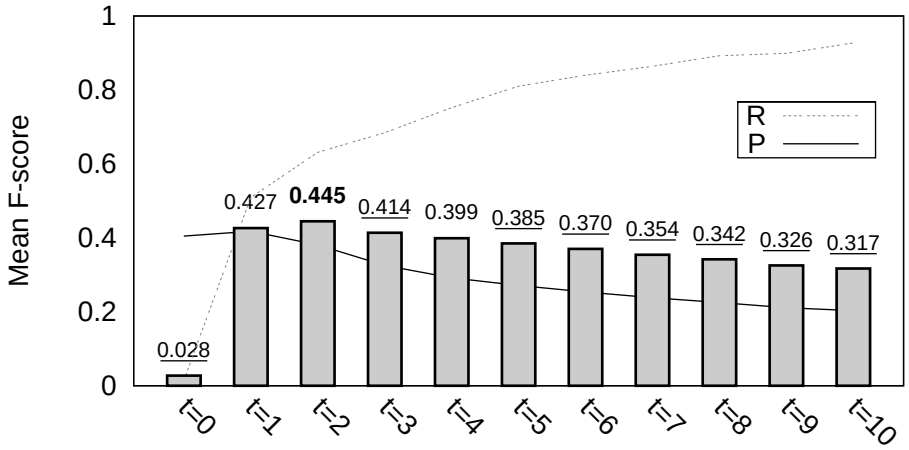


Figure 4.7: *Window size results for token-based model. The best score is in bold and those that are statistically distinguishable from the best are underlined.*

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 chapter, 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, demonstrating that as t is increased, recall goes up dramatically and precision goes down. Recall and precision are closest to being balanced at t=1.

4.3.4 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 the IsPair$_{deps}$ function defined in Figure 4.8, which counts all pairs of entity mentions

```
IsPair_deps : i, j, d
1  if sent(i) = sent(j) and count(dep-path-tokens(i, j)) ≤ d
2        return true
3  else
4        return false
```

Figure 4.8: *Function for GRI based on dependency path windows.*

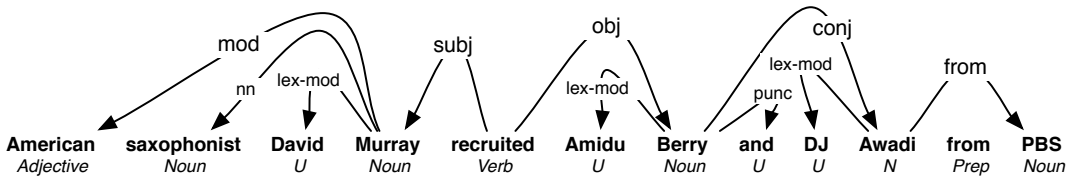
that 1) occur in the same sentence and 2) have d or fewer intervening token nodes on the shortest dependency path connecting the two entity mentions (the derivation of which is described in the next paragraph). This assumes access to a function (*dep-path-tokens*) that returns the set of token nodes on the dependency path connecting the two entity mentions indexed by i and j. While dependency paths have been successfully incorporated into supervised approaches to relation extraction, they have not been used for GRI.

For the current work, dependency paths are derived from syntactic parses obtained from the Minipar software. Minipar (Lin, 1998) is a broad-coverage parser based on an efficient message passing architecture with a lexicon derived from WordNet and a statistical ranking mechanism for selecting the best parse.[3] Minipar produces syntactic parse information in the form of typed grammatical relations including 1) the directional link from governors to their dependent lexical items and 2) grammatical relation types (e.g., *subject*, *object*). Chapter 3 and Appendix A contain further details of the pre-processing, which includes tokenisation, dependency parsing and the addition of the resulting governor-dependency relations to the XML representation of the documents. Figure 4.9(a) contains the Minipar parse of the example sentence from Figure 4.4. Dependency relations include, e.g. a *modifier* (mod) relation from governor noun "Murray" to dependent adjective "American", a *subject* (subj) relation from governor verb "recruited" to dependent noun "Murray", a *object* (obj) relation from "recruited" to dependent noun "Berry", a *prepositional modifier* (from) relation from governor noun "Awadi" to dependent noun "PBS".[4]

[3]Minipar achieves approximately 79% coverage of the dependency relationships in the SUSANNE corpus with 89% precision (Lin, 1998). The current evaluation only considers Minipar because the purpose here is to determine whether dependency information is useful for the GRI task, not to find the best dependency parser for the GRI task. For comparable systems, see e.g. Briscoe and Carroll (2006), de Marneffe et al. (2006).

[4]The *prepositional modifier* (from) relation from governor noun "Awadi" to dependent noun "PBS" is not actually a single grammatical relation in the Minipar output. It originally consists of two relations:

a) Dependency parse for example sentence



b) Shortest dependency paths between candidate entity mention pairs

1	American	$\xleftarrow{\textbf{mod}}$	David_Murray						
2	Amidu_Berry	$\xrightarrow{\textbf{conj}}$	Awadi						
3	Awadi	$\xrightarrow{\textbf{from}}$	PBS						
4	David_Murray	$\xleftarrow{\textbf{subj}}$	recruited	$\xrightarrow{\textbf{obj}}$	Amidu_Berry				
5	David_Murray	$\xleftarrow{\textbf{subj}}$	recruited	$\xrightarrow{\textbf{obj}}$	Awadi				
6	Amidu_Berry	$\xrightarrow{\textbf{conj}}$	Awadi	$\xrightarrow{\textbf{from}}$	PBS				
7	American	$\xleftarrow{\textbf{mod}}$	Murray	$\xleftarrow{\textbf{subj}}$	recruited	$\xrightarrow{\textbf{obj}}$	Amidu_Berry		
8	American	$\xleftarrow{\textbf{mod}}$	Murray	$\xleftarrow{\textbf{subj}}$	recruited	$\xrightarrow{\textbf{obj}}$	Awadi		
9	David_Murray	$\xleftarrow{\textbf{subj}}$	recruited	$\xrightarrow{\textbf{obj}}$	Berry	$\xrightarrow{\textbf{from}}$	PBS		
	David_Murray	$\xleftarrow{\textbf{subj}}$	recruited	$\xrightarrow{\textbf{obj}}$	Awadi	$\xrightarrow{\textbf{from}}$	PBS		
10	American	$\xleftarrow{\textbf{mod}}$	Murray	$\xleftarrow{\textbf{subj}}$	recruited	$\xrightarrow{\textbf{obj}}$	Berry	$\xrightarrow{\textbf{from}}$	PBS
	American	$\xleftarrow{\textbf{mod}}$	Murray	$\xleftarrow{\textbf{subj}}$	recruited	$\xrightarrow{\textbf{obj}}$	Awadi	$\xrightarrow{\textbf{from}}$	PBS

Figure 4.9: *Example dependency parse and dependency paths for all entity mention pairs.*

The shortest dependency paths between all candidate entity mention pairs are then extracted from the parse graph. Figure 4.9(b) contains dependency paths for the example sentence from Figure 4.4. Line 1, for example, contains the dependency path between "American" and "David Murray". This consists of a direct *modifier* (mod) relation with zero intervening word token nodes. Line 4, on the other hand, contains a dependency path (between "David Murray" and "Amidu Berry") that passes through 1 word token node ("recruited"). Line 5 contains a collapsed path that is a result of a post-processing operation over the Minipar output that passes governor-dependency relations along chains of conjoined tokens in the intervening context. Based on the parse

a *modifier* (mod) relation from governor "Awadi" to dependent "from" and a *preposition complement* (pcomp) relation from governor "from" to dependent "PBS". These are collapsed to a single relation as a post-processing step following Lin and Pantel (2001). This serves to connect the prepositional complement directly to the words modified by the preposition.

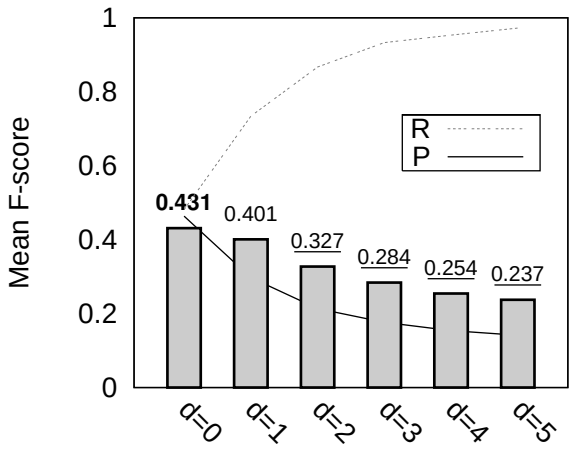


Figure 4.10: *Window size results for dependency-based model. The best score is in bold and those that are statistically distinguishable from the best are underlined.*

graph from Figure 4.9(a), the shortest path between "David Murray" and "Awadi" would actually pass through two word token nodes (instead of one) and include three relations: a *subject* (subj) relation between governor recruited and dependent "Murray", an *object* (obj) relation between recruited and "Berry" and a *conjunction* (conj) relation between "Berry" and "Awadi". The collapsing operation removes the *object* and *conjunction* relations, replacing them with a single *object* relation from governor "recruited" to dependent "Awadi". In cases like Lines 9 and 10 where the conjunction collapsing operation means there are multiple paths of the same length, the first path is chosen.

The example sentence and the relation-forming pairs extracted by the dependency path (Deps) model with various settings of the threshold d can be seen in Figure 4.4. The fifth and sixth rows correspond to models with d set to 0 and 1 respectively. The d=0 system does well in terms of precision on the example sentence, generating only one false positive entity mention pair (<Amidu_Berry,Awadi>). It does worse in terms of recall, missing two gold standard relation mentions (<David_Murray, Amidu_Berry> and <David_Murray,Amidu_Berry>). The d=1 system picks up the same false positive relation mention but achieves perfect recall on the example sentence.

Figure 4.10 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 (actual value printed above the bars) for different settings of d, which are spec-

ISPAIR$_{comb}$: i, j, t, d

1 **if** $sent(i) = sent(j)$ **and** ($count(intervening\text{-}tokens(i, j)) \leq t$
or $count(dep\text{-}path\text{-}tokens(i, j)) \leq d$)

2 **return** $true$

3 **else**

4 **return** $false$

Figure 4.11: *Function for GRI based on combined (token and dependency) windows.*

ified along the bottom of the horizontal axis. The best f-score is shown in bold and is achieved at d=0. Values that are statistically distinguishable (i.e., $p \leq 0.05$) 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, demonstrating that as d is increased, recall goes up dramatically while precision goes down. Recall and precision are closest to being balanced at d=0.

4.3.5 Combined Windows

Finally, the current work also introduces an entity mention co-occurrence model that combines token and dependency windows (Comb). It uses the ISPAIR$_{comb}$ function defined in Figure 4.11. This 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. This assumes access to two functions (*intervening-tokens* and *dep-path-tokens*) that are described in Sections 4.3.3 and 4.3.4 above.

An example sentence and the relation-forming pairs extracted by the combined (Comb) model with the intervening token threshold t set to 2 and the dependency path threshold d set to 0. The system does well in terms of precision on the example sentence, generating one false positive relation mention (<Amidu_Berry,Awadi>). It also does reasonably well in terms of recall, missing just one relation mention (<David_Murray,Awadi>).

Figure 4.12 contains joint optimisation results for the intervening token (t) and dependency path (d) thresholds on the news development data (ACE 2004). The shaded bars correspond to mean f-score (actual value printed above the bars) for different settings of t and d, which are specified along the bottom of the horizontal axis. The

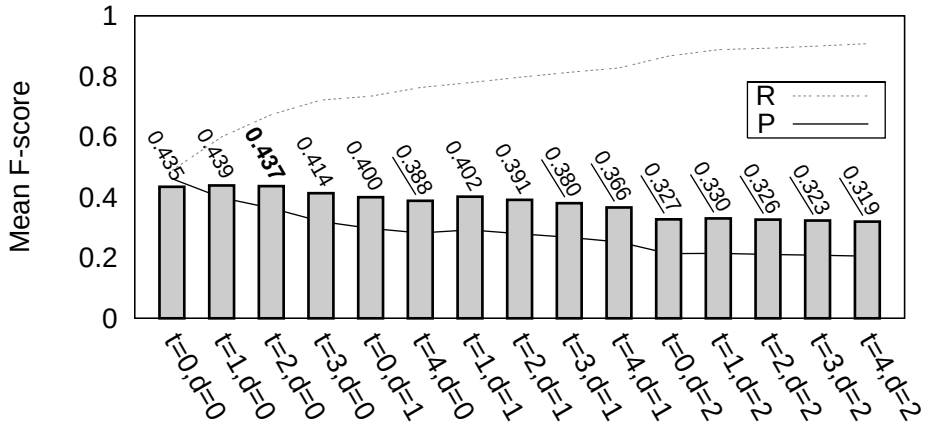


Figure 4.12: *Window size results for combined (token and dependency) model. The f-score of the system that achieves the highest mean rank is in bold and those that are statistically distinguishable from the best are underlined.*

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.[5] Values that are statistically distinguishable from the best (i.e., $p \leq 0.05$) are underlined. The results suggest a range of optimal settings with t ranging from 0 to 2 and d ranging from 0 to 1. The system with t=3 and d=0 is also statistically indistinguishable from the best. Recall (R) and precision (P) are plotted as dotted grey and solid black lines respectively.

4.4 Evaluation Experiments

4.4.1 Experiment 1: Model Comparison

4.4.1.1 Method

The first experiment compares the various window-based models for GRI. Specifically, it addresses the following question:

- *What window function is best for identifying relation mentions?*

This directly compares the intervening token, dependency path and combined models for entity mention co-occurrence. All models use window configurations optimised

[5]Note that the system with t=2 and d=0 is also better than the system with t=1 and d=0 in terms of recall, the prioritisation of which is discussed in Section 4.4.1.2 below.

	P	R	F_μ
Toks	0.291	0.510	0.342
Deps	**0.456**	0.392	**0.360**
Comb	0.277	**0.538**	0.332

Table 4.2: *Comparison of precision (P), recall (R) and f-score (F) results for token-based (Toks), dependency-based (Deps) and combined (Comb) systems on news test set. The best score in each column is in bold and those that are statistically distinguishable from the best are underlined.*

on the news development data (see Section 4.3). The intervening token model uses a threshold of t=2; the dependency path model uses a threshold of d=0; and the combined model uses thresholds of t=2 and d=0.

4.4.1.2 Results

Table 4.2 contains precision (P), recall (R) and f-score (F) results. Rows in the table correspond to the intervening token (Toks), dependency path (Deps) and combined (Comb) models. The best score for each evaluation measure is in bold. Systems that are statistically distinguishable from the best for the given measure (i.e., $p \leq 0.05$) are underlined. The highest f-score 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. The results suggest that the the Toks model with (t=2) contributes more in terms of recall, though the Deps model (with d=0) obtains higher precision.

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, the relation mining work discussed in Chapter 2 uses statistical measures of association such as pointwise mutual information, ϕ^2 and log likelihood ratio to estimate association strengths. Furthermore, the extrinsic evaluation in Chapter 6 follows the methodology of Filatova and Hatzivassiloglou (2004), who use similar models of atomic events

based on weighted pairs of co-occurring entities in the context of extractive summarisation. 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. This assumption is supported by the analysis in Section 4.5.4 and by the extrinsic evaluation experiments in Chapter 6.

4.4.2 Experiment 2: Comparison to Performance Bounds

4.4.2.1 Method

The second experiment evaluates the accuracy of the combined window-based model with respect to lower and upper bounds. It addresses the following questions:

- *Can GRI be improved using window-based models optimised on gold standard data?*
- *How does the optimised window-based model compare to human performance?*

This evaluates the contribution of the combined model with respect to a baseline approach from the related literature (see Section 4.3.1 above). It also compares to a human upper bound derived from the ACE double annotation (see Section 4.2.2). The window configuration uses thresholds of t=2 and d=0.

4.4.2.2 Results

Table 4.3 contains precision (P), recall (R) and f-score (F) results. Rows in the table correspond to the baseline model (Baseline), combined co-occurrence window model (Comb) and the human agreement (Human). The best score for each evaluation measure is in bold. Systems that are statistically distinguishable from the best for the given measure (i.e., $p \leq 0.05$) are underlined. The recall for the Baseline model is perfect because it counts all pairs of entity mentions that occur in the same sentence. This is equivalent to the GRI approach used by the Smith (2002) and Zhang et al. (2005) systems discussed in Chapter 2. The results in Table 4.3 demonstrate that the Comb model outperforms the Baseline model ($p = 0.0078$).

There is room for improvement with respect to the Human upper bound. The main difference is in terms of precision, where the Comb model performs far worse than the Human upper bound. However, while Comb recall is significantly worse than Human recall ($p = 0.0391$), the difference is not large. Furthermore, it should be

	P	R	F_μ
Baseline	0.110	**1.000**	0.195
Comb	**0.277**	0.538	**0.332**
Human	0.906	0.675	0.773

Table 4.3: *Precision (P), recall (R) and f-score (F) results for combined window-based system (Comb) with respect to baseline (Baseline) and human upper bound (Human) on news test set. The best score in each column is in bold and those that are statistically distinguishable from the best are underlined.*

noted that 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 pre-defined schema and, as shown in the analysis below (Section 4.5.2.1), 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.

4.4.3 Experiment 3: Integrating Long-Distance Relation Mentions

4.4.3.1 Method

The third experiment looks at the potential contribution of constrained long-distance relation mentions. It addresses the following question:

- *Is it possible to improve generic relation identification using filtering constraints (i.e., by requiring a verb or nominalisation in the intervening context)?*

This seeks to assess the potential impact of incorporating the atomic event identification from Filatova and Hatzivassiloglou (2003), which considers all pairs of entity mentions that occur in the same sentence but constrains them by requiring a verbal 'connector' as explained in Section 4.3.2. The hypothesis here is that the connector constraint may allow the model to incorporate some long-distance relation mentions with relatively high precision. The combined model again uses the optimised thresholds of t=2 and d=0.

	P	*R*	*F*
Event	0.050	0.392	0.083
Comb	**0.277**	**0.538**	**0.332**

Table 4.4: *Precision (P), recall (R) and f-score (F) results for combined window-based system (Comb) with respect to Filatova and Hatzivassiloglou (2004) atomic event system (Event) on news test set. The best score in each column is in bold and those that are statistically distinguishable from the best are underlined.*

4.4.3.2 Results

Table 4.4 contains precision (P), recall (R) and f-score (F) results. Rows in the table correspond to the atomic event model (Event) – considered here as a potential model of constrained long-distance relation mentions – and the combined window-based model (Comb). The best score for each evaluation measure is in bold and systems that are statistically distinguishable from the best for the given measure (i.e., $p \leq 0.05$) are underlined. The Comb model is the better in terms of overall f-score, giving an error rate reduction over the Event model with respect to the human upper bound of 36.1%. Because the Event model is constrained to detect relation mentions that are predicated by a verbal connector, it is not surprising that recall is not higher than the other systems. However, the very low precision suggests that the atomic event approach cannot be used as a high-precision metric to capture some long-distance relation mentions predicated by a verbal connector and improve the overall f-score. This question is also explored in the analysis below (Section 4.5.3), where the connector constraint is considered as a possible filter for improving the precision of smaller window functions.

4.4.4 Experiment 4: GRI Across Domains

4.4.4.1 Method

Finally, the fourth experiment addresses the claim of modification-free domain adaptation (i.e., that models achieve comparable accuracy when transferred, without modification of model parameters, across domains). It poses the following question:

- *Does model performance generalise across data sets and domains?*

Specifically, the performance of the various models are compared across the news domain and the biomedical domain. These models are optimised on the news development data (ACE 2004) and applied directly to the news (ACE 2005) and biomedical (BioInfer) test sets without modification. Results for the baseline and event models are also presented for comparison.

4.4.4.2 Results

Table 4.5 contains precision (P), recall (R) and f-score (F) results. Rows in the table correspond to the baseline model (Baseline), the atomic event model (Event), the intervening token model (Toks), the dependency path model (Deps) and the combined model (Comb). The best score for each evaluation measure is in bold and systems that are statistically distinguishable from the best (i.e., $p \leq 0.05$) are underlined. Table 4.5(a) repeats the results for the news domain test set (ACE 2005). Table 4.5(b) contains the results for the biomedical domain test set (BioInfer).

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 among entity mentions that occur in the same sentence. The biomedical result is consistent with the news result, however, in that Comb precision is significantly better than Baseline precision on both domains.

The result for the Event model is also consistent across domains. Precision is very low with respect to the other models (significantly worse than Toks, Deps and Comb at $p \leq 0.05$). This lends further support to the conclusion that the verbal connector constraint cannot be used as a high-precision metric to capture some long-distance relation mentions and improve f-score.

a) ACE 2005 (News Test Set)

	P	R	F_μ
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

b) BioInfer (Biomedical Test Set)

	P	R	F_μ
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**

Table 4.5: *Comparison of precision (P), recall (R) and f-score (F) results on news and biomedical test sets. Rows correspond to the baseline (Baseline), atomic event (Event), intervening tokens (Toks), dependency path (Deps) and combined (Comb) models. The best score in each column is in bold and those that are statistically distinguishable from the best are underlined.*

4.5 Analysis

4.5.1 Precision and Recall of Entity Pair Sub-Domains

Table 4.6 contains precision/recall scores for each entity pair sub-domain. Rows in Table 4.6(a) correspond to the entity pair sub-domains of the news test set where entity types include FACILITY/VEHICLE/WEAPON (F), GEOGRAPHICAL/POLITICAL/LOCATION (G), ORGANISATION (O) and PERSON (P). Rows in Table 4.6(b) correspond to the entity pair sub-domains of the biomedical test set where entity types include AMINO-ACID (A), SUBSTANCE (B), PROTEIN-COMPLEX (C), PROTEIN-FAMILY (F), NUCLEIC-ACID (N), INDIVIDUAL-PROTEIN (P), SOURCE (R) and PROTEIN-SUBSTRUCTURE (S). Columns correspond to the GRI models described in Section 4.3.

As mentioned above (Section 4.4.2.2), the baseline model accepts all possible entity pairs so achieves perfect recall. Thus, precision scores reflect the rate of true relation mentions in each entity pair sub-domain. As discussed in Section 4.4.3, 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 G-G and BioInfer P-F 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

a) ACE 2005 (News Test Set)

SD	Baseline	Event	Toks	Deps	Comb
F-G	0.171/1.000	0.027/0.100	0.609/0.700	0.813/0.650	0.560/0.700
F-P	0.196/1.000	0.146/0.556	0.310/0.250	0.615/0.222	0.364/0.333
G-G	0.067/1.000	0.006/0.057	0.216/0.862	0.158/0.517	0.177/0.874
G-O	0.049/1.000	0.020/0.265	0.194/0.618	0.169/0.588	0.115/0.618
G-P	0.125/1.000	0.081/0.522	0.327/0.323	0.688/0.264	0.347/0.383
O-P	0.115/1.000	0.048/0.319	0.289/0.471	0.627/0.353	0.286/0.496
P-P	0.050/1.000	0.019/0.295	0.091/0.344	0.120/0.148	0.087/0.361

b) BioInfer (Biomedical Test Set)

SD	Baseline	Event	Toks	Deps	Comb
A-N	0.191/1.000	0.169/0.651	0.333/0.140	0.300/0.070	0.300/0.140
N-N	0.119/1.000	0.132/0.531	0.102/0.265	0.085/0.204	0.090/0.286
P-C	0.415/1.000	0.319/0.462	0.744/0.246	0.642/0.277	0.662/0.377
P-F	0.270/1.000	0.060/0.088	0.755/0.720	0.872/0.637	0.752/0.819
P-P	0.278/1.000	0.197/0.269	0.407/0.463	0.380/0.465	0.397/0.602
P-S	0.362/1.000	0.242/0.400	0.750/0.508	0.689/0.238	0.763/0.569
R-B	0.242/1.000	0.182/0.529	0.600/0.375	0.523/0.221	0.533/0.385

Table 4.6: *Precision/recall on entity pair sub-domains for news and biomedical test sets.*

2005 G-G sub-domain tend to be geographical part-of relations where the two entity mentions are adjacent (e.g., the relation between the G entity mention "Peoria" and the G entity mention "Illinois" in the fragment "Peoria, Illinois"). And, true relation mentions in the BioInfer P-F sub-domain tend to be appositives (e.g., the relation between the P entity mention "cofilin" and the F entity mention "actin-binding protein" in the fragment "cofilin, a ubiquitous actin-binding protein") or nominal modifiers (e.g., the relation between the F entity mention "cyclin-dependent kinase inhibitors" and the P entity mention "p57" in the fragment "the cyclin-dependent kinase inhibitors (CKIs) p27 and p57").

With respect to the intervening tokens (Toks), dependency path (Deps) and combined (Comb) models on the ACE 2005 test set (Table 4.6(a)), Toks is generally better in terms of recall while Deps is better in terms of precision. Recall is generally highest for the combined model, achieving large improvements with respect to the Toks and

Deps models on the F-P sub-domain. Precision for the combined model is generally in between Toks and Deps precision though closer to the lower of the two.

On the BioInfer test set (Table 4.6(b)), the combined model is again best in terms of recall, achieving large improvements with respect to the Toks and Deps models on the P-C, P-F and P-P sub-domains. Toks, rather than Deps, is generally better in terms of precision, which is probably due in part to lower parse accuracy (see Section 4.5.2 below). However, all systems actually do better in terms of precision on the BioInfer test data (as compared to the ACE 2005 test data). This can be attributed at least in part to the much higher rate of true relation mentions in the BioInfer test set (mean across sub-domains of 26.8%) with respect to the ACE 2005 test set (11.0%).

4.5.2 Error Analysis

This section contains an analysis that aims to characterise the types of errors made by the combined (Comb) GRI system. For each entity pair sub-domain, ten instances are chosen randomly from the set of erroneously classified instances. These are manually inspected to determine the error types. The analysis is presented in two parts: 1) Section 4.5.2.1 contains a breakdown of error types for false positive (FP) classifications where the system posits a relation mention that the annotators do not and 2) Section 4.5.2.2 contains a breakdown of error types for false negative (FN) classifications where the system misses a relation mention that is present in the annotation.

4.5.2.1 False Positives

In Tables 4.7(a) and 4.7(b), the columns correspond to entity pair sub-domains as described in Section 4.5.1 above.[6] The first row in the table corresponds to the count of FPs among the random sample of ten erroneously classified instances, the second row corresponds to the percentage of FPs where the number of intervening tokens is within the threshold (t=2) and the third row corresponds to the percentage of FPs where the number of token nodes on the dependency path is within the threshold (d=0). The fourth row in Table 4.7(a) corresponds to the percentage of FPs where one or both entity mentions are pronouns while the fourth row in Table 4.7(b) corresponds to the percentage of FPs where the entity mentions are embedded (i.e., the begin and end tokens of one entity mention span are within the begin and end tokens of the other entity mention span). The rest of the rows contain the breakdown of error types, which are

[6]The P-C column of Table 4.7(b) is all NAs because there were no FPs among the error sample.

described with examples in the remainder of this section. These are grouped into three categories: 1) unequivocal system errors, 2) system errors where a relation is more or less implicit given the context of the sentence and 3) annotation errors where the system is actually correct.

Bad Parse instances are FP errors that are due to a parse error. This is exemplified by the relation mention predicted in the BioInfer data between the SOURCE (R) entity mention "RVS161" and the SUBSTANCE (B) entity mention "actin cytoskeleton" (due to an erroneous *conjunction* dependency relation between "RVS161" and "cytoskeleton") in the following sentence:

> "Mutations in RVS161 and RVS167, the two yeast amphiphysin homologs, cause very similar growth phenotypes, a depolarized actin cytoskeleton, and a defect in the internalization step of endocytosis."

This type of error is rare, occurring once in each of the ACE G-P, ACE P-P and BioInfer R-B sub-domains. This suggests that parse errors do not have a large effect on the precision of the combined model though an improved parser would lead to slightly higher precision in the sub-domains mentioned.

Model Noise instances are FP errors that are due to over-generation by the GRI model. This is exemplified by the relation mention that is predicted in the ACE data between the FACILITY/VEHICLE/WEAPON (F) entity mention "air force one" and the PERSON (P) entity mention "him" in the following sentence:[7]

> "the president greeting a number of dignitaries as he gets ready here to board air force one on his way home."

This type of error is somewhat common. It is difficult to address due to the nature of the task, which adopts shallow approaches to favour generic solutions over the best possible performance on a given domain or task. Section 4.5.3 below considers some possible indicator features to improve precision by filtering FP errors.

Comparison instances are FP errors where there is arguably a relation mention that makes an explicit comparison. This is exemplified by the relation mention that is predicted in the ACE data between the PERSON (P) entity mention "Gul" and the PERSON (P) entity mention "Erdogan" in the following sentence:

> "Unlike the soft-spoken Gul, Erdogan has a reputation as a fighter."

[7] As discussed in Chapter 3, the ACE news data is sourced from both newswire and broadcast news. Text snippets from the broadcast news transcripts are presented as-is, without any capitalisation and including any disfluencies.

a) ACE 2005 (News Test Set)

	F-G	F-P	G-G	G-O	G-P	O-P	P-P
N	6	7	10	10	3	8	8
Toks Accept	100%	100%	80%	90%	67%	88%	75%
Deps Accept	0%	14%	70%	40%	33%	13%	38%
Pronominal Entities	100%	71%	20%	20%	67%	63%	50%
System Error	0%	29%	0%	10%	67%	13%	0%
Bad Parse	–	–	–	–	*33%*	–	–
Model Noise	–	*29%*	–	*10%*	*33%*	*13%*	–
Implicit Relation	100%	71%	100%	90%	33%	75%	100%
Comparison	–	–	–	–	–	–	*13%*
Conjunction	–	–	*70%*	*10%*	–	*13%*	*38%*
Figurative	*20%*	–	–	*10%*	–	–	–
Formulaic	–	–	–	*70%*	–	–	–
Future/Past	*20%*	–	–	–	–	–	–
Identity	–	–	*10%*	–	–	–	*38%*
Inferable	*60%*	*71%*	*20%*	–	*33%*	*63%*	*13%*
True Rel (Annot Err)	0%	0%	0%	0%	0%	13%	0%

b) BioInfer (Biomedical Test Set)

	A-N	N-N	P-C	P-F	P-P	P-S	R-B
N	2	6	0	4	7	4	2
Toks Accept	100%	33%	NA	100%	43%	100%	50%
Deps Accept	0%	67%	NA	0%	71%	50%	100%
Embedded Entities	0%	67%	NA	0%	43%	25%	50%
System Error	0%	0%	NA	25%	0%	25%	50%
Bad Parse	–	–	NA	–	–	–	50%
Model Noise	–	–	NA	*25%*	–	25%	–
Implicit Relation	50%	100%	NA	75%	100%	75%	50%
Conjunction	–	*83%*	NA	–	*43%*	*25%*	–
Identity	–	–	NA	–	*29%*	–	–
Inferable	*50%*	*17%*	NA	*75%*	*29%*	*50%*	*50%*
Other	–	–	NA	–	*0%*	–	–
True Rel (Annot Err)	50%	0%	NA	0%	0%	0%	0%

Table 4.7: *Breakdown of FP error types for combined token- and dependency-based model on news and biomedical test sets.*

This type of error only occurs just once in the ACE P-P sub-domain so is not a substantial problem. This could actually be argued to be an explicit relation mention, but it does not fit into the relation type schema that was used to guide the annotation task.

Conjunction instances are FP errors where the entity mentions are in a list or otherwise conjoined, which sometimes indicates implicit relation concerning similarity. This is exemplified by the relation mention that is predicted in the BioInfer data between the NUCLEIC-ACID (N) entity mention "tropomyosin" and the NUCLEIC-ACID (N) entity mention "Abp1p" in the following sentence:

> "A null mutation of the actin gene (ACT1) is lethal, but null mutations in the tropomyosin (TPM1), fimbrin (SAC6), Abp1p (ABP1), and capping protein (CAP1 and CAP2) genes have relatively mild or no effects."

This is the second most common type of FP error in both the ACE and BioInfer data, accounting respectively for 19% and 25% of the total FP errors (mean across sub-domains). This occurs primarily in symmetric domains (i.e., where both entity mentions have the same type). Section 4.5.3 investigates the use of conjunction indicator features as a filter for reducing FP errors among system predictions.

Figurative instances are FP errors where the relation mention is embedded in figurative language such as metaphors. This is exemplified by the relation mention that is predicted in the ACE data between the GEOGRAPHICAL/POLITICAL/LOCATION (G) entity mention "we" and the FACILITY/VEHICLE/WEAPON (F) entity mention "bridge" in the following sentence:

> "We are not facing that kind of situation but we will cross that bridge when we come to it."

This type of error is rare occurring once in each of the ACE F-G[8] and ACE G-O sub-domains, which suggests that this error does not affect the model much. While a lexicon of figurative speech may help to filter these instances, the rarity of the error means it is probably not worth the additional machinery.

Formulaic instances are FP errors where the relation mention is embedded in a standardised phrase structure such as newspaper bylines. This is exemplified by the relation mention that is predicted in the ACE data between the GEOGRAPHICAL/POLITICAL/LOCATION (G) entity mention "ANKARA" and the ORGANISATION (O) entity mention "AP" in the following fragments:

[8]Note that the entity mention "we" is annotated as GEOGRAPHICAL/POLITICAL/LOCATION (G) because it is taken from a quote in which U.S. Ambassador John Negroponte is replying to a question about American foreign policy and is referring to the country's government when he says we.

> "ANKARA, Turkey (AP)"

This type of error was only found to occur in the G-O domain but accounts for 70% of the errors there. This could be addressed by developing a pre-processing system to identify formulaic phrases. In fact, this is how Hasegawa et al. (2004) handle by-lines in their experimental data from the New York Times. However, formulaic phrases are domain-specific so this is not a generic solution that would apply across e.g. domains and tasks.

Future/Past instances are FP errors where a relation mention is evident from the text but it is a relation that held in the past or is expected to hold in the future. This is exemplified by the relation mention that is predicted in the ACE data between the FACILITY/VEHICLE/WEAPON (F) entity mention "itself" and the GEOGRAPHICAL/POLITICAL/LOCATION (G) entity mention "hawaii" in the following sentence:

> "his son was on one of the ships that escorted a carrier. although this came home by itself from hawaii."

This type of error only occurs once in the ACE F-G sub-domain so is not a substantial problem. While temporal modelling may help to filter these instances, the rarity of the error means it is probably not worth the additional machinery.

Identity instances are FP errors where the entity mentions are coreferent (i.e., they refer to the same underlying object). This is exemplified by the relation mention that is predicted in the BioInfer data between the INDIVIDUAL-PROTEIN (P) entity mention "chick actin-depolymerizing factor" and the INDIVIDUAL-PROTEIN (P) entity mention "ADF" in the following sentence:

> "Two cDNAs, isolated from a Xenopus laevis embryonic library, encode proteins of 168 amino acids, both of which are 77% identical to chick cofilin and 66% identical to chick actin-depolymerizing factor (ADF), two structurally and functionally related proteins."

This type of error is rare, occurring once in the ACE G-G sub-domain and twice in the BioInfer P-P sub-domain. Cases like the one above could be easily addressed using existing methods for abbreviation detection (e.g., Schwartz and Hearst, 2003; Torii et al., 2006). However, because the error is rare, it may not be worth the additional machinery.

Inferable instances are FP errors where a relation can be inferred from the context. This is exemplified by the relation mention that is predicted in the BioInfer data between the AMINO-ACID (A) entity mention "RAD52 proteins" and the NUCLEIC-ACID (N) entity mention "RAD51" in the following sentence:

> "Because other researchers have shown that the RAD51 and RAD52 proteins interact, RAD51 on a high copy number plasmid was tested and found to suppress the rad52-20 allele, but RAD 54, 55 and 57 did not suppress."

Here, there is a clear statement of interaction between the first mention of entity mention "RAD51" and the entity mention "RAD52 proteins" while the relation between "RAD52 proteins" and the second mention of "RAD51" is inferable from the context but not marked in the gold standard. This type of FP error is the most common in both the ACE and BioInfer data, accounting respectively for 37% and 45% of the total FP errors (mean across sub-domains). The intervening token threshold model is generally more prone to this type of error (100% and 78% of errors respectively on ACE and BioInfer) than the dependency path model (0% and 33%). Thus, it could be addressed by using the dependency path model exclusively. However, the development results in Section 4.3 above demonstrate that this also results in a large loss in recall.

True Rel instances are FP errors where a relation mention should have been posited in the gold standard annotation. This is exemplified by the relation mention missing from the BioInfer data between the NUCLEIC-ACID (N) entity mention "histone" and the AMINO-ACID (A) entity mention "H4" in the following sentence:

> "The histone H4 and histone H2b genes encode 10% of the total H4 and H2b mRNA."

Annotator errors are very rare occurring only once each in the FP samples for the ACE and BioInfer data sets.

4.5.2.2 False Negatives

In Tables 4.8(a) and 4.8(b), the columns correspond to entity pair sub-domains as described in Section 4.5.1 above.[9] The first row of the table corresponds to the count of FNs among the random sample of ten erroneously classified instances. The second row in Table 4.8(a) corresponds to the percentage of FNs where one or both entity mentions are pronouns while the second row in Table 4.8(b) corresponds to the percentage of FNs where the entity mentions are embedded (i.e., the begin and end tokens of one entity mention span are within the begin and end tokens of the other entity mention span). No embedded entity mentions are actually recorded in Table 4.8(b), however the 0 counts are left in to illustrate the contrast with embedded entity mention counts for the FP instances (Table 4.7(b)). The rest of the rows contain the breakdown of FN

[9] The G-G and G-O columns of Table 4.8(a) are all NAs because there were no FNs among the error sample.

a) ACE 2005 (News Test Set)

	F-G	F-P	G-G	G-O	G-P	O-P	P-P
N	4	3	0	0	7	2	2
Pronominal Entities	25%	33%	NA	NA	29%	50%	100%
System Error	75%	67%	NA	NA	86%	100%	100%
Bad Parse	*25%*	–	NA	NA	–	–	–
Thresh Error	*50%*	*67%*	NA	NA	*86%*	*100%*	*100%*
Quest Rel (Annot Err)	25%	33%	NA	NA	14%	0%	0%
Future/Past	–	–	NA	NA	*14%*	–	–
World Knowledge	*25%*	*33%*	NA	NA	–	–	–

b) BioInfer (Biomedical Test Set)

	A-N	N-N	P-C	P-F	P-P	P-S	R-B
N	8	4	10	6	3	6	8
Embedded Entities	0%	0%	0%	0%	0%	0%	0%
System Error	100%	100%	100%	83%	67%	83%	67%
Bad Parse	*13%*	*25%*	–	*33%*	*33%*	–	–
Thresh Error	*88%*	*75%*	*100%*	*50%*	*33%*	*83%*	*67%*
Quest Rel (Annot Err)	0%	0%	0%	17%	33%	17%	33%
World Knowledge	–	–	–	*17%*	*33%*	*17%*	*33%*

Table 4.8: *Breakdown of FN error types for combined token- and dependency-based model on news and biomedical test sets.*

error types which are described with examples in the remainder of this section. These are grouped into two categories: 1) unequivocal system errors and 2) questionable relation mentions (Quest Rel) where the validity of the annotated relation mention may be called into question (i.e., the error is arguably in the annotation, not in the system prediction).

Bad Parse instances are FN errors that are due to a parse error. This is exemplified by the relation mention that is missed in the BioInfer data between the INDIVIDUAL-PROTEIN (P) entity mention "cofilin" and the PROTEIN-FAMILY (F) entity mention "actin-binding protein" (due to the bad parse output missing an *appositive* governor-dependency relation between "cofilin" and "protein") in the following sentence:

> "Here we identify a pathway for the regulation of cofilin, a ubiquitous actin-binding protein that is essential for effective depolymerization of actin filaments."

This type of error occurs only once in the ACE data but is the second most common in the BioInfer data, accounting for 15% of the total FN errors (mean across sub-domains). This suggests that the Minipar parser does not perform as well on biomedical text as it does on news text. Another parser that has been developed or modified for biomedical text processing may do better (e.g., Hara et al., 2005; Briscoe et al., 2006) but evaluating modification-free domain adaptation without a specialised parser is a stronger test condition.

Thresh Error instances are FN errors that are due to coverage limitations of the tuned models (i.e., the number of intervening tokens is greater than two and the number of token nodes on the dependency path is greater than zero). This is exemplified by the relation mention missed in the ACE data between the ORGANISATION (O) entity mention "tyco" and the PERSON (P) entity mention "dennis kozlowski" in the following fragment:[10]

> "tyco's ceo and president dennis kozlowski"

This type of FN error is the most common in both the ACE and BioInfer data, accounting respectively for 81% and 71% of the total FN errors (mean across sub-domains). These recall errors can be addressed simply by increasing the intervening token or dependency path thresholds. However, the development results in Section 4.3 demonstrate that this also results in a large decrease in precision. Thus, as mentioned in Section 4.5.2.1 above with respect to FP errors due to model noise, further improvements require methods to filter false positives from high recall systems. Some possible filtering approaches are examined in the following section (Section 4.5.3).

Future/Past instances are FN errors where a relation mention is evident from the text but it is a relation that held in the past or is expected to hold in the future. This exemplified by the relation mention that is missed between the PERSON (P) entity mention "own" and the GEOGRAPHICAL/POLITICAL/LOCATION (G) entity mention "iraq" in the following sentence:[11]

> "chalabi staged his own rally yesterday to support his bid to become the next leader of iraq."

[10] The dependency path for this fragment has one intervening node ("president") and two governor-dependency relations. The first is a *subject* relation between "tyco" and "president" (which should technically be a *possessive* relation but the wrong type does not change the number of nodes). The second is a *person* relation between "president" and "kozlowski". The number of intervening word tokens is four.

[11] Arguably, the relation mention should be between "chalabi" and "iraq". However, even if this were the case, the relation mention would still be missed by the system and, more to the point, it would still be a possible future relation as opposed to a relation mention that is true in the context of the sentence and document.

This type of error is very rare, occurring only once in the ACE G-P sub-domain. It is interesting, though, that this kind of temporal confusion occurs both among the false positive errors and among the false negative errors, suggesting that the annotators did not have a clear idea of how to the interaction between time and relations should have been treated.

World Knowledge instances are FN errors where a relation is not clearly stated in the sentence but is implicit, requiring reasoning or external domain/world knowledge. This is exemplified by the USER-OWNER-INVENTOR-MANUFACTURE relation mention that is missed between the GEOGRAPHICAL/POLITICAL/LOCATION entity mention "We" and the FACILITY/VEHICLE/WEAPON entity mention "Australian embassy" in the following sentence:

> "We have quite a substantial security presence at the Australian embassy in Riyadh ... ."

Here, it would be necessary to know that "We" is used by an Australian official to refer to the country's government and that it is the Australian government that owns and uses the Australian embassy. This type of FN error is the second most common in the ACE data and the third most common in the BioInfer data, accounting respectively for 12% and 14% of the total FN errors (mean across sub-domains). It is interesting that implicit errors occur both among the false positives and the false negatives, suggesting some confusion among annotators as to whether these should be marked.

4.5.3 Feature-Based Filtering of FP Errors

Due to the nature of the window-based models, recall can be improved simply by increasing token or dependency windows (illustrated in Figures 4.7 and 4.10 above). However, this also results in lower precision. Thus, improved f-scores require methods to filter false positives from high recall window-based models. Table 4.9 contains correlation (phi coefficient) scores that compare various binary indicator features with a binary variable indicating whether the instance constitutes a true relation mention according to the annotation. Since the focus is on filtering false positives, only entity mention pairs that are predicted to be relation mentions by a system are considered. The rows in Table 4.9 correspond to the various models. The columns correspond to the various indicator features which include presence of an event connector word in intervening token context (CN_t), presence of a conjunction/disjunction in intervening token context (CJ_t), presence of an event connector word on the dependency path

a) ACE 2005 (News Test Set)

	CN_t	CJ_t	CN_d	CJ_d	US_d	EN_d	MB_1	MB_2
Baseline	***-0.290***	***-0.124***	**-0.322**	-0.050	***-0.118***	***-0.116***	NA	NA
Toks	***-0.253***	-0.041	**-0.415**	NA	-0.094	-0.099	NA	NA
Deps	NA	***-0.296***	NA	***-0.293***	NA	NA	NA	NA
Comb	***-0.195***	-0.076	**-0.360**	***-0.175***	-0.068	-0.055	NA	NA

b) BioInfer (Biomedical Test Set)

	CN_t	CJ_t	CN_d	CJ_d	US_d	EN_d	MB_1	MB_2
Baseline	***-0.186***	***-0.285***	***-0.194***	***-0.108***	***-0.164***	***-0.143***	-0.084	-0.043
Toks	0.083	***-0.209***	-0.038	***-0.207***	***0.100***	-0.083	NA	***-0.285***
Deps	-0.009	***-0.133***	NA	***-0.200***	NA	NA	NA	**-0.310**
Comb	0.051	***-0.208***	-0.004	***-0.242***	***-0.102***	-0.092	NA	***-0.233***

Table 4.9: *Phi coefficient correlation analysis comparing a true relation mention indicator feature to various indicator features for filtering false positives errors from GRI output.*

(CN_d), presence of a conjunction/disjunction on the dependency path (CJ_d), presence of a unspecified governor-dependency relation on dependency path (US_d),[12] presence of an empty node on the dependency path (EN_d),[13] whether the tokens of one entity mention are a subset of the tokens of the other entity mention (MB_1) and whether one entity mention is actually contained within the other in the text (MB_2).

Following conventions in the literature for effect size of the phi coefficient (e.g., Cohen, 1988; Coolican, 2004), 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. NA values indicate cases where the phi coefficient is undefined due to one or both of the variables having zero variance. The table suggests that the presence of a connector word on the dependency path (CN_d) would be the strongest filter of false positive instances for ACE 2005 with medium negative correlations of -0.322, -0.415 and -0.360 respectively for the Baseline, Toks

[12]Unspecified governor-dependency relations are artifacts of Minipar dependency parsing where the relation type could not be determined by Minipar.

[13]Empty nodes are artifacts of Minipar dependency parsing. Empty nodes are nodes that do not correspond to a specific word token in the input.

and Comb models.[14] This effect is also evident, though smaller, for the Baseline model on the BioInfer test set. The table also suggests that the embedding of entity mentions (MB_2) would be the strongest filter of false positive instances for BioInfer with small to medium negative correlations of -0.285, -0.310 and -0.233 respectively for the Toks, Deps and Comb models.[15]

Thus, using the existence of event connector words as a filter should increase precision on ACE 2005 and either have no effect or increase precision slightly on BioInfer. By contrast, using embedding of entity mentions as a filter should increase precision on BioInfer and have no significant effect on ACE 2005. Considering only the combined model (Comb), a number of other indicator features are also candidate filters (i.e., in order of effect size: CJ_d, CJ_t, CN_t, US_d) though the effect size is smaller and detailed experiments would be necessary to see if any corresponding reduction in recall is detrimental to f-scores and to determine the effect of dependencies between various features (e.g., CJ_t and CJ_d). These results in combination with the development results in Section 4.3 suggest that it would be worth looking at higher recall systems (e.g., Deps with d=1) that use the filtering constraints discussed here to improve precision.

4.5.4 Comparison of Ranking Methods

Another possible method for improving precision would be to incorporate methods from the literature for ranking entity mention pairs using statistical measures of association. Section 2.4.1 describes several such methods including pair probability (Pr), log-likelihood (G^2), ϕ^2, and pointwise mutual information (PMI). Table 4.10 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. NA values indicate cases where the point-biserial coefficient is undefined due to one or both of the variables having zero variance. The table suggests that a threshold filtering low values of

[14]The correlation result suggesting that the presence of a connector word on the dependency path (CN_d) is somewhat contrary to the previous results in which the Event model achieved very low precision scores (Sections 4.5.1 and 4.4.3 above). This suggests that a refined Event model based on dependency paths may do better than the Filatova and Hatzivassiloglou (2004) Event model evaluated above.

[15]The correlation with the entity mention embedding indicator features (i.e., MB_1 and MB_2) is not measurable in the ACE data due to the the mapping of embedded relation mentions described in Section 3.3.1.2.

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

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.

4.6 Summary and Future Work

This chapter presented experiments with window-based models for the generic relation identification (GRI) task. 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. And the use of gold standard annotation in both domains allowed detailed analysis, exploring the behaviour of the various models.

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 (including the summarisation task addressed in Chapter 6) generally incorporate a mechanism for ranking identified relation mentions. Correlation analysis supported this prioritisation of recall, suggesting that ranking metrics can be used as a weak indicator of true/false relation status. Based on the output of the Comb model, pointwise mutual information (PMI) demonstrated a medium effect on the news data while log-likelihood (G^2) obtained a small effect that was more consistent across domains.

Experimental results also showed that optimisation of the window-based models leads to an improvement over a baseline approach from the literature that accepts all entity mention pairs occurring in the same sentence. Comparison to human performance suggests that there is room for considerable improvement of all models, though the performance of human annotators is a very strong upper bound due to the fact that they performed a highly constrained task of marking relation mentions according to specific guidelines and a relation type schema as opposed to the generic task that is being evaluated.

The optimised window-based model was also compared to a model for atomic events (Event) that is similar to the window-based relation models, but also requires a verbal connector word in the intervening context. This model was significantly worse than the Comb model in terms of f-score. The precision of the Event model was very low, refuting the hypothesis that this could be used as a constrained, high-precision metric for identifying some long-distance relation mentions. Furthermore, related correlation analysis suggested that false positive filters based on the presence of intervening connectors would serve to improve precision of models with smaller windows (e.g., intervening tokens threshold of t=2 or dependency path threshold of d=0).

Experiments and analysis suggest that GRI accuracy is comparable when applying the newswire-optimised models directly to the biomedical domain. In both domains the best recall is achieved by the Comb model and the f-score is at least as good as the next best model (in the biomedical domain, the Comb f-score is actually significantly better

than the Deps f-score). One not unexpected difference 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).

The error analysis also demonstrated 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). 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 obtaining reliable rankings e.g. for weighting relations for entity sketches (as discussed in Chapter 3) or for representing the conceptual content of a sentence for extractive summarisation (as in the extrinsic evaluation in Chapter 6).

Future work will look at improving the precision of GRI models. In conjunction with development experiments, correlation analysis of indicator features for filtering false positives suggests that constrained high recall systems may lead to further accuracy improvements for GRI. For example, the dependency path model might be improved by including entity mention pairs that have one intervening node on the dependency path but requiring that the intervening node be a verbal connector. Other possible constraints include a requirement that the intervening node not be a conjunction or that the two entity mentions not be embedded.

Finally, accuracy may also be improved through refinement of the dependency model. Greenwood and Stevenson (2007) compare several pattern representations on a relation identification task, where the goal of the system is to identify pairs of entity mentions that are part of the same event according to a gold standard corpus. They find that the top-ranked patterns for a highly constrained representation have precision between 0.8 and 0.9 and that more expressive representations identify patterns with precision that is consistently in the range of 0.2 to 0.4, degrading slowly with increased recall, the most expressive representation achieving recall as high as 0.8. The more constrained approaches could be explored as a means of improving the accuracy and efficiency of GRI. As another example, Banko and Etzioni (2008) list eight simplified syntactic patterns that cover 95% of the binary relation mentions in an IE corpus of 500 sentences. Among these patterns, they find that the pattern consisting of two entity's with an intervening verb covers 37.8% of binary relation mentions.

Chapter 5

Generic Relation Characterisation

Experiments are reported that address the generic relation characterisation task, comparing similarity models that are parametrised by feature set and dimensionality reduction technique. A novel feature set is introduced for the task based on syntactic features from governor-dependency parses. Comparison of dimensionality reduction techniques shows that a similarity model based on latent Dirichlet analysis (LDA) – a probabilistic generative approach – successfully incorporates a larger and more interdependent feature set than an unreduced model and a model based on singular value decomposition (SVD). LDA offers as much as a 34.5% reduction in the error rate when compared to SVD. And, while not always significant, it achieves higher f-scores than other approaches on five out of six evaluation settings. Taken together with the superior interpretability of the probabilistic generative approach, this motivates the use of LDA in applications of generic relation extraction. Furthermore, these models are shown for the first time to achieve comparable accuracy when transferred across domains.

5.1 Introduction

This chapter addresses the generic relation characterisation (GRC) task, where the goal is to induce a partition (or clustering) over relation-forming entity mention pairs (or relation mentions) that groups them by relation type.[1] Figure 5.1 contains an overview of the GRC task, which is split into three main sub-tasks. The input is a collection of natural language documents with entity mentions and relation-forming pairs identified.[2]

[1]Full GRC includes a cluster labelling step, which is addressed in the relation discovery literature (e.g., Hasegawa et al., 2004; Chen et al., 2005) and discussed in Chapter 2. For the purposes of this chapter, we focus on the primary modelling task with respect to entity mention pair clustering and refer to this sub-task when using the term GRC.

[2]For the evaluation here, the input includes gold standard relation-forming entity mention pairs as discussed in Chapter 3. The ACE data input only includes relation-forming pairs over entities mentions

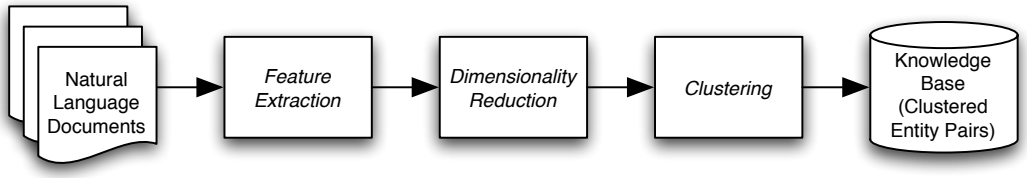


Figure 5.1: *Overview of GRC clustering sub-tasks.*

The first sub-task has the goal of identifying features from the context of the entity mention pairs that are indicative of relation type and outputs a feature-based representation of the pairs. The second sub-task introduces a novel step into the GRC pipeline, where the feature space is optionally transformed using dimensionality reduction techniques. Finally, in the third sub-task, the actual clustering is performed. The output consists of a partition over relation-forming pairs that groups them by relation type.

As discussed in Chapter 2, previous GRC work has largely failed to use standardised data or evaluation measures, making it difficult to compare approaches. This chapter employs a principled framework for evaluation (introduced in Chapter 3) that makes use of gold standard relation extraction data to optimise and evaluate GRC models. News data from the ACE shared tasks (described in Chapter 3) is used for development and for testing on a held-out evaluation set in the same domain. The presence of double annotation in the ACE 2005 data makes it possible to compute a human upper bound for the GRC task. Additionally, biomedical data from BioInfer (also described in Chapter 3) is also used, allowing assessment of model consistency across application domains.

In terms of modelling, previous GRC work has relied extensively on various lexical and shallow syntactic features of the context surrounding entity mention pairs (e.g., word tokens, part-of-speech, chunk phrase information). As discussed in Chapter 2, Zhang et al. (2005) introduced a clustering model based on tree kernels, derived from parse trees obtained from a phrase structure parser (Collins, 1999). However, previous approaches do not exploit dependency parsing, which provides typed governor-dependency relations between word tokens as opposed to the syntactic constituency information provided by phrase structure parsing (e.g. Mel'čuk, 1987; de Marneffe et al., 2006). Dependency parse information has been successfully incorporated into supervised (including rule-based) approaches to relation extraction (e.g., Bunescu and

that are named, prenominal or pronominal as discussed at the end of Section 3.3.1.2.

Mooney, 2007; Fundel et al., 2007), so it is a natural extension to GRC. The current work incorporates features based on dependency parsing information in a framework that also uses features from the intervening context. Another shortcoming of previous models of GRC is that they rely on direct matching of features for computing similarity, which fails to identify similarities between features with different surface strings but similar underlying (or latent) semantics. The current work also introduces the use of dimensionality reduction and compares two common approaches used in the language processing and information retrieval literature.

A detailed comparison of previous GRC work can be found in Chapter 2. This chapter begins with a description of the setup for experimental evaluation in Section 5.2. Next, Section 5.3 contains a specification of the models that are compared here. Section 5.4 contains experimental results and discussion. Finally, Section 5.5 contains a detailed analysis of the experimental results.

5.2 The Task: Experimental Setup

5.2.1 GRC Framework

This chapter explores models for GRC that are based on novel feature sets derived from dependency parsing and transformation of the resulting feature space using dimensionality reduction techniques from the language processing literature. These two primary modelling parameters correspond to the first two sub-tasks in Figure 5.1 and will be discussed in detail in the modelling section below (Section 5.3). This section addresses the other parameters that are held constant for the experiments presented here. These include 1) the approach to model order selection (i.e., determining the number of clusters), 2) the measure for quantifying similarity between feature vectors and 3) the clustering algorithm itself.

5.2.1.1 Model Order Selection

Model order selection is the task of determining the number of clusters in a data set. This has been variously treated in the GRC literature. Chen et al. (2005) and Chen et al. (2006) use approaches from the literature for automatic model order selection (see Chapter 2). Hasegawa et al. (2004), on the other hand, use hierarchical agglomerative clustering based on the fact that the task is exploratory in nature and the number of clusters or desired granularity would not be known in advance and is moreover de-

pendent on the application scenario. Furthermore, while related gold standards exist that could be used to tune automatic order selection techniques, these gold standards themselves are based on arbitrary decisions about the depth and breadth of the relation type schema. This point is illustrated through a quick comparison of the ACE (LDC, 2004c, 2005b) and BioInfer (Pyysalo et al., 2007) data sets (Cf. full schemas in Appendix B). Where ACE 2004 has a relatively simple relation type schema consisting of a hierarchy with 2 levels and 22 leaf nodes (2 and 18 respectively for ACE 2005), the BioInfer schema consists of a hierarchy that is 6 deep in places and has a total of 68 leaf nodes.

For the evaluation in this chapter, the gold standard number of clusters is used. This is motivated first by the fact that this allows the dimensionality reduction models to be explicitly and efficiently tuned to represent the density and skew of the cluster distribution in the development data as discussed in the experimental sections below. Second, this is motivated by the fact that the output of dimensionality reduction rather than the subsequent clustering output is used for the extrinsic evaluation (see Chapter 6). While this weakens the claim that the GRC approach learns the relation type schema for the development data, the resulting tuned models can still be considered to learn the relation type schema for unseen data. Particularly when one considers that it is possible to either 1) follow Hasegawa et al. (2004) in outputting the dendrogram from the hierarchical clustering rather than a partition based on the hierarchical clustering or 2) perform automatic model order selection based on approaches from the literature as demonstrated in the full GRE system output in Section 3.2.4.

5.2.1.2 Measuring Similarity

Cosine is commonly used in the literature to compute similarities between *tf*idf*-weighted feature vectors. This is defined as:

$$Cosine(p,q) = \frac{\sum_i p_i q_i}{\sqrt{\sum_j p_j^2}\sqrt{\sum_k q_k^2}} \tag{5.1}$$

where p and q are feature vectors. In the current work, cosine similarity is used for unreduced feature and SVD representations (described below in Section 5.3.2).

The choice of measure for quantifying similarity for probabilistic models such as LDA (described below in Section 5.3.2) is based on previous work (Hachey, 2006), where Kullback-Leibler (KL) divergence was compared with symmetrised KL divergence and Jensen-Shannon (JS) divergence on the GRC task. Experiments showed

that KL performed significantly better than symmetrised KL and was indistinguishable from JS. Due to the fact that the mean KL score was slightly higher than JS and KL is more efficient, KL is used for the experiments here. KL divergence is defined as:

$$KL(p||q) = \sum_i p_i \log \frac{p_i}{q_i} \tag{5.2}$$

where p and q are probability distributions over the same event space. In information-theoretic terms, KL divergence is the average number of bits wasted by encoding events from a distribution p with a code based on distribution q.

Technically, the divergence measures are dissimilarity measures as higher values correspond to larger differences (smaller similarities) between distributions. However, they can be converted to increasing measures of similarity through various transformations. The choice here is also motivated by previous work (Hachey, 2006), which found the simple approach from Lee (1999) ($Sim_{Lee}(p,q) = C - KL(p||q)$, where C is a free parameter to be tuned) to outperform a related approach from Dagan et al. (1997) ($Sim_{Dagan}(p,q) = 10^{-\beta KL(p||q)}$). Also, the divergence-to-similarity conversion metric was found to interact with the dimensionality, motivating joint optimisation of the two parameters. Optimised values for the experiments here are reported below in Section 5.3.2.

5.2.1.3 Clustering Techniques

Clustering is performed with the CLUTO software[3] and the technique used is identical across models. Hierarchical agglomerative clustering is used for comparability with the original relation discovery work of Hasegawa et al. (2004) (discussed in Chapter 2 and Section 5.2.1.1 above).

One way to view the clustering problem is as an optimisation process where an optimal clustering is chosen with respect to a global criterion function over the entire solution (Zhao and Karypis, 2004). Global criterion functions include:

I_1 - a function that maximises the sum of pairwise similarities between instances assigned to each cluster;

I_2 - an internal function that maximises the similarity between each instance and the centroid of the cluster it is assigned to;

[3] http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview

$\mathcal{E}_1$ - an external function that minimises the similarity between the centroid vector of each cluster and the centroid vector of the entire collection; and

$\mathcal{H}_1$ - a combined (internal and external) function that consists of the ratio of I_1 over $\mathcal{E}_1$.

Another way to view the agglomerative clustering problem is as a greedy optimisation process where an optimal clustering is chosen based on a series of greedy decisions concerning which two clusters to merge next (Zhao and Karypis, 2005; Manning et al., 2008). Local criterion functions include:

single-link - a function that measures similarity between clusters by their two most similar members;

complete-link - a function that measures similarity between clusters by their two least similar members; and

group average - a function that measures similarity between clusters as the average similarity between their members.

In preliminary experiments on the development data, the I_2, $\mathcal{H}_1$ and $\mathcal{H}_2$ criterion functions were found to outperform single link, complete link and group average on the development data in terms of accuracy. The experiments reported below use I_2, which had accuracy comparable to that of $\mathcal{H}_1$ and $\mathcal{H}_2$ but is superior in terms of computational complexity (Cf., Zhao and Karypis, 2004).

5.2.2 Data and Evaluation

The evaluation uses news data from the Automatic Content Extraction (ACE) 2004 and 2005 shared tasks and biomedical data derived from the BioInfer corpus (see Chapter 3 for details of data sets and preparation). 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 held-out biomedical test set. As discussed in Chapter 3, evaluation of clustering is in terms of both the 1-to-1 f-score ($F_{1:1}$) and the pairwise f-score (F_{pw}). These are both calculated with respect to the partition defined by the gold standard relation annotation (Equations 3.15 and 3.17 in Chapter 3). Differences are tested for statistical significance using paired Wilcoxon signed ranks tests across the entity pair sub-domains (also described in Chapter 3).

a) 1-to-1 f-score ($F_{1:1}$)

	P	R	F
Human 1	0.947	0.935	0.941
Human 2	0.991	0.990	0.991
Mean Human	0.969	0.963	0.966

b) Pairwise f-score (F_{pw})

	P	R	F
Human 1	0.913	0.884	0.897
Human 2	0.979	0.990	0.984
Mean Human	0.946	0.937	0.941

Table 5.1: *Precision (P), recall (R) and f-score (F) results for human annotators against adjudicated gold standard.*

Another aspect of the evaluation here is the introduction of an upper bound based on human agreement. The ACE 2005 data includes markup from two human annotators and a final adjudicated version of the markup, which makes it possible to compute inter-annotator 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 optimised one-to-one mapping based on maximal character overlap between entity mention strings LDC (2004a). Given this mapping, it is possible to align relation-forming pairs and extract partitions for each annotator and the adjudicated gold standard. The partitions derived from the individual annotators are then evaluated against the gold standard in the same way as the systems.

Table 5.1 contains precision (P), recall (R) and f-score (F) results for the individual human annotators when compared to the final adjudicated data set. The first two rows contain the individual annotator results and the bottom row contains the mean of the two individual annotators. The mean agreement is 0.966 and 0.941 in terms of $F_{1:1}$ and F_{pw} respectively.

5.3 Models

The focus of the experiments in this chapter includes 1) novel feature sets derived from dependency parsing and 2) transformation of the resulting feature space using dimensionality reduction techniques. These correspond to the first two steps of the GRC clustering sub-task in Figure 5.1. In the remainder of this section, these two modelling concerns are discussed as experimental parameters.

5.3.1 Features Beyond Intervening Words

The first experimental parameter explored here is the feature representation for GRC. In this section, several feature sets are described for representing relation types. Figure 5.2 contains an example sentence and the various feature representations extracted for the relation-forming entity mention pairs. The first row contains the example sentence where entity mention starts and ends are marked with square brackets and the entity type is indicated by the superscript text to the right of the opening bracket. In the remaining rows of the table, the feature type is specified in the first column, the entity mention pair is given in the second and the features extracted are given in the second column. The feature sets are described in detail in the following subsections.

5.3.1.1 Intervening Word Features

The Intervening Word (W) features are based on the word tokens that occur in the intervening context between two relation-forming entity mentions. Stop words are removed using a publicly available list distributed with the Infomap NLP software.[4] Then intervening tokens are stemmed using the Porter stemming algorithm (Porter, 1980). All remaining non-stop tokens are kept as intervening word features. This feature space is equivalent to that used by Hasegawa et al. (2004) and Chen et al. (2005).[5] Chen et al. also experiment with word tokens from the sentence context before the first entity mention and after the second entity mention, though these are not found to improve scores.

An example sentence and the associated Intervening Word features for each relation-forming pair can be seen in Figure 5.2. Intervening Word features do well at capturing relevant information for verbal relation mentions like the BUSINESS relation mention between "David Murray" and "Amidu Barry", which is described by the intervening verb "recruited" (whose stem is "recruit"). However, Intervening Word features do poorly in capturing relevant information that is implicitly present in the underlying syntax or semantics. For example, *saxophonist* is the only feature extracted for the CITIZEN-RESIDENT-RELIGION-ETHNICITY relation mention between "American" and "David Murray". And no features are extracted for the MEMBER-OF-GROUP relation mention between "Awadi" and "PBS" or the CAUSAL relation mention be-

[4] http://infomap-nlp.sourceforge.net/

[5] Actually, the Intervening Words approach here may differ slightly from related work. Hasegawa et al. (2004) do not use stemming and do not say how their stop list is compiled. Chen et al. (2005) do not say whether they perform stemming or stopping.

a) ACE 2004 (News Development Set)

Example Sentence	[^{place} American] saxophonist [^{person} David Murray] recruited [^{person} Amidu Barry] and DJ [^{person} Awadi] from [^{organisation} PBS].	
Intervening Word	<American,David_Murray>	{*saxophonist*}
	<David_Murray,Amidu_Barry>	{*recruit*}
	<David_Murray,Awadi>	{*recruit, amidu, barri, dj*}
	<Amidu_Barry,PBS>	{*dj, awadi*}
	<Awadi,PBS>	{}
Entity Word	<American,David_Murray>	{}
	<David_Murray,Amidu_Barry>	{}
	<David_Murray,Awadi>	{}
	<Amidu_Barry,PBS>	{}
	<Awadi,PBS>	{}
Dependency Path	<American,David_Murray>	{*r_mod*}
	<David_Murray,Amidu_Barry>	{*r_subj, w_recruit, r_obj*}
	<David_Murray,Awadi>	{*r_subj, w_recruit, r_obj*}
	<Amidu_Barry,PBS>	{*r_conj, w_awadi, r_from*}
	<Awadi,PBS>	{*r_from*}

b) BioInfer (Biomedical Test Set)

Example Sentence	[^{gene} Cdc3+] encodes [^{protein} profilin], an [^{protein} actin-monomer]-binding protein.	
Intervening Word	<cdc3,profilin>	{*encod*}
	<profilin,actin-monomer>	{}
Entity Word	<cdc3,profilin>	{*profilin*}
	<profilin,actin-monomer>	{*profilin, actin, -monom*}
Dependency Path	<cdc3,profilin>	{*r_subj, w_encod, r_obj*}
	<profilin,actin-monomer>	{*r_appo, w_protein, r_mod, w_bind, r_lex-mod*}

Figure 5.2: *Example sentences with gold standard relation-forming entity mention pairs and corresponding feature representations for various feature sets.*

tween "profilin" and "actin-monomer". Furthermore, Intervening Word features often produce noise even when they produce relevant features, as in the *barri* and *dj* features extracted for the BUSINESS relation mention between "David Murray" and "Awadi".

5.3.1.2 Entity Word Features

The Entity Word (E) features are based on the word tokens that occur in the entity mention phrases themselves. First, a simple heuristic method is used to determine whether an entity mention is nominal. This uses part-of-speech tags to count the number of proper noun tokens (C_{pn}) and the total number of tokens (C_{ttl}) in a given entity mention phrase. Entity mentions are considered to be nominal if the proportion of proper noun tokens (C_{pn}/C_{ttl}) is less than 0.75. Non-nominal entity mentions are not used to generate Entity Word features. Next, stop words are removed from the set of entity mention word tokens using a list of function words, function word parts-of-speech and number tags. Function word lists are sourced from the per-class word frequency lists from the British National Corpus (BNC) web page.[6] Lists for conjunctions, determiners, pronouns, prepositions and interjections were used. Parts-of-speech from Minipar are used to identify multi-word prepositions. Finally, numeric tokens are identified using *num* relations from Minipar dependency parses. All remaining non-stop tokens are kept as entity mention word features. This is comparable to the Entity Word features introduced by Chen et al. (2005) in related work.

The specific Entity Word features used here are based on preliminary experiments on the ACE development data, where it was found that using only non-stop word tokens from nominal entity mentions was more beneficial than using word tokens from entity mentions of any mention type. The intuition is that nominal (including non-named prenominal) entity mentions may sometimes describe the relation type (e.g., "brother", "lawyer") while named entity mentions are not likely to describe the relation type. Consider the following fragment:

"the [^{person} Bush] [^{organisation} cabinet]"

The EMPLOY-EXECUTIVE relation mention between "Bush" and "cabinet" is described in part by the fact that a cabinet is something that a person can appoint and head (and in part by the underlying syntax). Furthermore, Entity Word features also contribute a level of coreference resolution, in that they make it more likely for entity mentions containing the same word tokens to be assigned to the same relation type cluster. This is

[6] BNC frequency lists were accessed on 28 November 2006 from `http://ucrel.lancs.ac.uk/bncfreq/flists.html`.

often beneficial, though can also cause errors. Consider the CAUSAL relation mention between "profilin" and "actin-monomer" in Figure 5.2(b) and the PART-OF relation mention between "pollen profilin isoform" and "ZmPRO1" in the following fragment:

"the [^{protein} pollen profilin isoform], [^{protein} ZmPRO1]"

The fact that the feature sets for both relation mentions contain the entity word "profilin" could lead to the two entity mention pairs being wrongly placed in the same relation type cluster.

As concerns the running examples, the Entity Word features include the *profilin* feature for the relation mention between "Cdc3+" and "profilin" in Figure 5.2(b), which has a gold standard relation type of IS-A. They do not include any features for "Cdc3+" (or any of the entity mentions in Figure 5.2(a)) because it is not identified as a nominal entity mention.

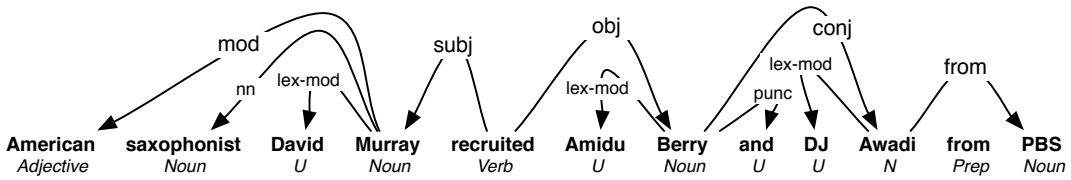
5.3.1.3 Dependency Path Features

The Dependency Path (D) features are based on words and grammatical relations from a dependency parser. This is a novel approach to representing relation type information for GRC clustering, which extracts features from the shortest dependency path connecting the two entity mentions (derivation described in the next paragraph). This is motivated by the fact that some relation mentions are described by the underlying syntax rather than the words themselves. Consider the CITIZEN-RESIDENT-RELIGION-ETHNICITY relation mention between "American" and "David Murray" in Figure 5.2(a). This relation type is not completely evident from the entity mentions or surrounding words alone, but it is evident given the fact that the underlying governor-dependency structure contains a modifier relation mention between the words "Murray" and "American". While dependency parse information has proved successful in supervised approaches to relation extraction (e.g., Bunescu and Mooney, 2007; Fundel et al., 2007), it has not been used for GRC.

As described in Chapter 4, dependency paths are derived from syntactic parses obtained from Minipar (Lin, 1998). Minipar produces syntactic parse information in the form of typed grammatical relations including 1) the directional link from governors to their dependent lexical items and 2) grammatical relation types (e.g., *subject*, *object*).[7] Figure 5.3 repeats the Minipar parse of the example news data sentence from Figure

[7]Section 4.3.4 contains further details of dependency parsing. Section 3.3 contains details of preprocessing, including tokenisation and dependency parsing. And, Appendix A contains details of the XML document type used here for encoding relation extraction data and linguistic annotation.

a) Dependency parse for example sentence



b) Shortest dependency paths between relation-forming entity mention pairs

1	American	$\xleftarrow{\textbf{mod}}$	David_Murray			
2	David_Murray	$\xleftarrow{\textbf{subj}}$	recruited	$\xrightarrow{\textbf{obj}}$	Amidu_Barry	
3	David_Murray	$\xleftarrow{\textbf{subj}}$	recruited	$\xrightarrow{\textbf{obj}}$	Awadi	
4	Amidu_Barry	$\xleftarrow{\textbf{conj}}$	Awadi	$\xrightarrow{\textbf{from}}$	PBS	
5	Awadi	$\xrightarrow{\textbf{from}}$	PBS			

Figure 5.3: *Example dependency parse and dependency paths for relation-forming entity mention pairs.*

5.2(a) and the associated dependency paths. In Chapter 4, dependency paths were used without governor-dependency relation type information. Here, on the other hand, type information is used to derive features for clustering. This is done by extracting relation types and word tokens from the dependency path. For the CITIZEN-RESIDENT-RELIGION-ETHNICITY relation mention between "American" and "David Murray" in Figure 5.2, for example, the dependency features include a dependency relation of type *modifier* (r_mod), which is extracted from the dependency path shown on Line 1 of Figure 5.3(b). This addresses the problem with the Intervening Word features, for which only the word "saxophonist" is extracted, which does not help to describe the relation type. Further examples of Dependency Path features can be seen in Figure 5.2 where dependency relations are prefixed with "r_" and words are prefixed with "w_".

5.3.2 Dimensionality Reduction

The second experimental parameter explored here is dimensionality reduction. Dimensionality reduction is a means of inferring latent structure in distributional data which has been argued to create models of semantic similarity that are more linguistic in nature (e.g., see the Landauer et al. (1998) discussion of LSA and synonym tests). Three approaches are compared here: 1) a baseline approach where no dimensionality

reduction is performed, 2) singular value decomposition (SVD) and 3) latent Dirichlet allocation (LDA). These are described in the rest of this section.

5.3.2.1 Unreduced Feature Space

The first approach here is a baseline approach where no dimensionality reduction is performed. Here, feature vectors are extracted for each relation mention and weighted using *tf*idf*, which is calculated as:

$$w(i,j) = \sqrt{tf_{i,j} * \log\left(\frac{N+1}{df_i}\right)} \tag{5.3}$$

where $tf_{i,j}$ is the number of times feature i occurs in the context of relation-forming entity mention pair j and df_i is the number of relation-forming pair contexts in which feature i occurs. As specified in Section 5.2.1.2 above, cosine is used to measure the similarity between feature vectors in the unreduced feature space.

5.3.2.2 SVD-Reduced Feature Space

The first dimensionality reduction technique employed in the current work is singular value decomposition (SVD), a linear algebraic least squares method (Eckart and Young, 1936). SVD has proved successful for related work in information retrieval and language processing. For example, Berry et al. (1994) describe an SVD-based approach to information retrieval that they term latent semantic indexing. And Landauer et al. (1998) describe the application of SVD-based latent semantic approaches to various cognitive modelling tasks including modelling synonymy, sorting words by word senses and modelling semantic priming. Where $X_{r\times f}$ is a relation-by-feature (R×F) matrix,[8] SVD performs a decomposition of X into the product of three matrices with n latent semantic dimensions:

$$X_{r\times f} = R_{r\times n} S_{n\times n} (F_{f\times n})^T$$

In the resulting decomposition, the R and F matrices represent relation mentions and features in the new space and S is a diagonal matrix of singular values in decreasing order. These are generally sorted by decreasing magnitude of the singular values.

This is illustrated in Figure 5.4. The box labelled X represents the original $r \times t$ matrix (with r rows representing relation mentions and f columns representing features). The boxes labelled R, S and F represent the full detail of the decomposition.

[8] Here, the input matrix to SVD is composed of the *tf*idf*-weighted feature vectors described in Section 5.3.2.1.

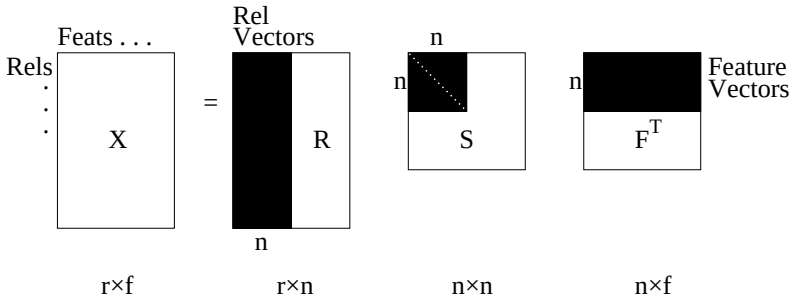


Figure 5.4: *Matrix visualisation of singular value decomposition (SVD).*

The shaded areas of these boxes represent the areas in the matrices corresponding to the n highest singular values. Thus, taking the product $R_{r\times n}S_{n\times n}(F_{f\times n})^T$ over the first n columns gives the best least squares approximation of the original matrix X by a matrix of rank n, i.e. a reduction of the original matrix to n dimensions. Similarly, a rank n representation of relation mentions can be derived by rescaling the first n columns of R with the first n singular values ($R_{r\times n}S_{n\times n}$).

The implementation used here is from the Python LinearAlgebra module, which provides interfaces to the LAPACK libraries in FORTRAN (Anderson et al., 1999). LAPACK is a suite of routines for linear algebra problems that includes various efficient solutions to singular value problems. SVD is performed using an implementation of Cuppen's divide and conquer algorithm to find the eigenvalues and the eigenvectors (Rutter, 1994). The model contains one free parameter: the dimensionality of the reduced space n. This is tuned on the ACE 2004 news development data. Values compared during tuning include the constant values $\{5, 10, 15, 25, 50, 75, 100\}$ and values specified as a proportion of the size f of the unreduced feature space $\{0.05f, 0.10f, 0.25f, 0.50f, 0.75f, 0.90f, 0.95f, 0.97f, 0.99f\}$ (rounded to the nearest whole number). Tuned n values used in the remaining experiments are given in the first data column (SVD n) of Table 5.2, where the rows correspond to the different combinations of intervening word (W), entity word (E) and dependency path(D) features.

The resulting rank n rescaled vectors ($R_{r\times n}S_{n\times n}$) are used as a latent semantic representation of relation mentions. As specified in Section 5.2.1.2, cosine is used to measure the similarity between two vectors in the SVD-reduced feature space.

Features	SVD n	LDA T	LDA C	LDA β	LDA α
W	15	$0.97F$	12	0.0001	$50/T$
D	$0.10f$	$0.99F$	20	0.001	10
WE	5	$0.99F$	15	$200/F$	$50/T$
WD	$0.10f$	$0.75F$	15	0.001	$50/T$
ED	$0.05f$	$0.90F$	20	0.0001	$50/T$
WED	5	$0.97F$	8	0.0001	$50/T$

Table 5.2: *Tuned SVD and LDA parameter values for various feature combinations based on intervening words (*W*), entity words (*E*) and dependency paths (*D*).*

5.3.2.3 LDA-Reduced Feature Space

While SVD has proved successful, its representation of words and documents (or relations) as points in a Euclidean space is not easy to interpret, motivating more recent work on analogous probabilistic models of latent semantic information. In early work, Hofmann (2001) introduced a generative probabilistic version of latent semantic analysis that models each word in a document as a sample from a mixture model. It does not, however, provide a model at the document (or relation) level. Latent Dirichlet allocation (LDA) addresses this by representing documents as random mixtures over latent topics (Blei et al., 2003). LDA has a clear probabilistic generative interpretation making its output easy to understand and making it easy to embed in larger applications. Thus, it is explored in the current work as an alternative to SVD.

Here, LDA is used to model the contribution of different topics to a relation mention by treating each topic as a probability distribution over features, where a relation mention is a probabilistic mixture of topics. Where T is the number of topics, the probability of the ith feature is written as:

$$P(f_i) = \sum_{j=1}^{T} P(f_i|z_i = j)P(z_i = j) \tag{5.4}$$

where z_i is a latent variable indicating the topic from which feature f_i is drawn, $P(f_i|z_i = j)$ is the probability of drawing feature f_i under topic j and $P(z_i = j)$ is the probability of topic j for the current relation mention. Intuitively, $P(\mathbf{f}|\mathbf{z})$ indicates which features are important to a topic and $P(\mathbf{z})$ is the prevalence of those topics for a given relation mention (Griffiths and Steyvers, 2004).

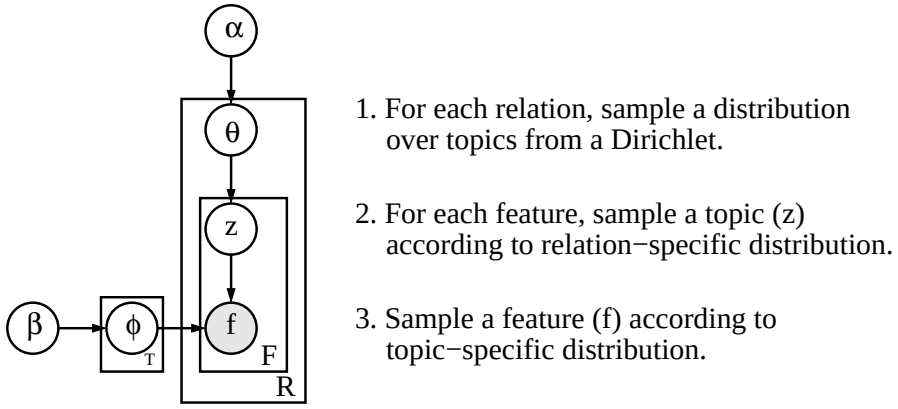


Figure 5.5: *Graphical representation of latent Dirichlet allocation (LDA).*

Figure 5.5 contains a graphical representation of the LDA model in plate notation, where nodes represent variables (shaded nodes corresponding to observed variables and non-shaded nodes corresponding to latent variables), arrows represent conditional dependencies and plates (boxes) represent repeated sampling steps (the variable in the lower right hand corner of plates corresponding to the number of repetitions). The numbered text on the right side of the figure describes the generative process. In the figure, ϕ represents the multinomial distributions over features for each topic, i.e. $P(f_i|z_i = j) = \phi_{f_i}^{(z_i=j)}$. And θ represents the multinomial distributions over topics for each relation mention, i.e. $P(z_i = j) = \theta_{z_i=j}^{(r_k)}$. R, F and T are the total number of relation mentions, features and topics respectively.

In its generative mode, the LDA model first chooses a topic distribution from a Dirichlet ($\theta \sim Dir(\alpha)$). Then, it samples a topic $z_i = j$ from the mention-specific multinomial distribution $\theta^{(r_k)}$. Finally, it samples a feature f_i from the topic-specific multinomial distribution $\phi^{(z_i=j)}$. The version of LDA used here also incorporates a Dirichlet prior over the multinomial distributions for features ($\phi \sim Dir(\beta)$), which was introduced by Blei et al. (2003) to address sparsity. The model here follows Griffiths and Steyvers (2004) in assuming symmetric Dirichlet priors with a single value each for β and α. These hyperparameters determine the nature of the priors, where values over one indicate a preference for multinomials that are closer to uniform and values under one indicate a preference for multinomials that are sparser (Goldwater and Griffiths, 2007).

The choice of the Dirichlet distribution as a prior is explained by its conjugacy to the multinomial distribution, meaning that if a multinomial distribution is endowed with a Dirichlet prior then the posterior will also be a Dirichlet. This allows efficient

estimation of the joint distribution over features and topics $P(\mathbf{f},\mathbf{z}) = P(\mathbf{f}|\mathbf{z})P(\mathbf{z})$ by integrating ϕ and θ out of the equations for $P(\mathbf{f}|\mathbf{z})$ and $P(\mathbf{z})$. Griffiths and Steyvers describe how this is performed to obtain a set of samples from the posterior distributions $P(\mathbf{z}|\mathbf{w})$ using Gibbs sampling, an approximate iterative method for sampling from complex distributions (Gilks et al., 1996). Given these samples from the posterior, a predictive distribution over topics θ can be estimated as:

$$\hat{\theta}_j^{(r)} = \frac{n_j^{(r)} + \alpha}{n_\cdot^{(r)} + T\alpha} \tag{5.5}$$

where $n_j^{(r)}$ is the number of times a word from relation mention r has been assigned to topic j.

The implementation of Gibbs sampling for LDA used here is provided in Mark Steyvers and Tom Griffiths' Matlab Topic Modeling Toolbox.[9] The resulting similarity model contains three free parameters: the number of topics T, the two hyperparameters (β and α) and the constant C for divergence-to-similarity conversion. These are tuned on the ACE 2004 news development data. First, T and C are tuned jointly with possible values of T being the same as for SVD $\{5,10,15,25,50,75,100,0.05F,0.10F,$ $0.25F,0.50F,0.75F,0.90F,0.95F,0.97F,0.99F\}$ and possible values of C including $\{0,5,8,10,12,15,20\}$. Next, values compared for β include $\{0.0000001,0.000001,$ $0.00001,0.0001,0.001,0.01,0.1,1,10,100,200/F\}$.[10] Lastly, values compared for α include $\{0.0001,0.001,0.01,0.1,1,10,100,1000,50/T\}$.[11] Tuned values used in the remaining experiments are given in the last four columns of Table 5.2, where the rows correspond to the different combinations of intervening word (W), entity word (E) and dependency path(D) features..[12]

The resulting topic distributions $\hat{\theta}^{(r)}$ are used as a latent semantic representation of relation mentions r. As specified in Section 5.2.1.2, $C - KL(p||q)$ is used to measure the similarity between two vectors in the LDA-reduced feature space.

[9] `http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm`.

[10] $200/F$ is the default value for β in the Topic Modeling Toolbox software.

[11] $50/T$ is the default value for α suggested by Steyvers and Griffiths (2007).

[12] Tuned values used for the final experiments here are based on optimisation using the 1-to-1 f-score ($F_{1:1}$. Results are similar for the pairwise f-score (F_{pw}).

5.4 Evaluation Experiments

5.4.1 Experiment 1: Model Comparison

5.4.1.1 Method

The first experiment compares the various combinations of feature sets and dimensionality reduction techniques. Specifically, it addresses the following questions:

1. *Which dimensionality reduction technique is best?*

2. *Which combination of features is best?*

Models are compared on the ACE 2004 news development data using the dimensionality reduction parameters from the tuning experiments described above (Table 5.2).

5.4.1.2 Results

Table 5.3 contains mean f-score (across entity pair sub-domains) results for six different feature sets: intervening words (W), dependency path (D), W combined with entity words (WE), W and D combined (WD), D combined with entity words (ED) and all three combined (WED). Rows in the table correspond to the unreduced clustering model (Cl:None), the SVD reduced model (Cl:SVD) and the LDA reduced model (Cl:LDA). The best score for each feature set is in bold. Systems that are statistically distinguishable from the best for the given feature set (i.e., $p \leq 0.05$) are underlined. The table suggests that the LDA-reduced similarity models are generally better, obtaining statistically better mean f-scores on many feature sets (i.e., WD and ED for $F_{1:1}$ and D, WE, WD, ED and WED for F_{pw}).

Interestingly, the best unreduced and SVD-reduced models are obtained with the intervening word (W) features while the best LDA-reduced models are obtained when feature combinations including word and dependency path features (i.e., the WD and WED feature sets) are used. This suggests that dimensionality reduction with LDA helps to incorporate a larger and more interdependent feature set. This may be explained by the LDA hyperparameters, which control the impact of sparsity (as described in Section 5.3.2.3 above). Specifically, nearly all of the tuned similarity models specified in Table 5.2 above have small β values of 0.0001 or 0.001. This can be expected to result in a fine-grained decomposition into topics that address specific relation types. Such a specification is not possible with SVD and indeed the tuned

b) 1-to-1 F-Score ($F_{1:1}$)

	W	D	WE	WD	ED	WED
Cl:None	0.664	0.572	0.614	0.563	0.584	0.591
Cl:SVD	**0.674**	0.568	**0.657**	0.624	0.567	0.609
Cl:LDA	0.638	**0.668**	0.653	**0.706**	**0.675**	**0.683**

a) Pairwise F-Score (F_{pw})

	W	D	WE	WD	ED	WED
Cl:None	0.628	0.466	0.543	0.469	0.471	0.486
Cl:SVD	**0.651**	0.489	0.612	0.534	0.473	0.505
Cl:LDA	0.597	**0.643**	**0.664**	**0.697**	**0.662**	**0.676**

Table 5.3: *Comparison of f-scores for dimensionality reduction techniques on news development set. Rows correspond to the unreduced (Cl:None), SVD-reduced (Cl:SVD) and LDA-reduced (Cl:LDA) clustering models. Columns correspond to the different combinations of intervening word (*W*), entity word (*E*) and dependency path (*D*) features. The best score in each column is in bold and those that are statistically distinguishable from the best are underlined.*

a) 1-to-1 f-score ($F_{1:1}$)

Hypothesis	p
W<WE	0.2344
W<D	0.1473
W<WD	0.0711
W<ED	0.0467*
W<WED	0.0463*
WE<ED	0.2344
WE<WED	0.2344
WE<D	0.1875
WE<WD	0.1179
D<ED	0.4170
D<WED	0.3375
D<WD	0.1473
ED<WD	0.5000
ED<WED	0.3937
WD<WED	0.2008

b) Pairwise f-score (F_{pw})

Hypothesis	p
W<D	0.0467*
W<WE	0.0391*
W<WD	0.0180*
W<ED	0.0180*
W<WED	0.0178*
D<WE	0.4688
D<ED	0.3375
D<WED	0.1042
D<WD	0.1036
ED<WE	0.4688
ED<WD	0.2008
ED<WED	0.0295*
WE<WED	0.1548
WE<WD	0.0234*
WD<WED	0.3937

Table 5.4: *Partial ranking of feature combinations for LDA-reduced similarity model based on Wilcoxon p values.*

parameters for SVD are very different, with SVD-reduced models performing better with few topics (see Table 5.2 above).

Table 5.3 suggests that the best performance for the LDA-reduced model is achieved with the WD feature set. However, a more detailed look suggests that the WED feature set may be preferable. Table 5.4 contains a partial ranking of feature sets for the LDA-reduced similarity model based on p values for paired Wilcoxon signed ranks tests, where statistically significant values ($p \leq 0.05$) are marked with a star. While the WD<WED p values are not significant, they suggest that WED may be preferable. Based on these results, the similarity models in the rest of the experiments use the combined WED feature set and LDA dimensionality reduction unless otherwise specified.

5.4.2 Experiment 2: Comparison to Performance Bounds

5.4.2.1 Method

The second experiment evaluates the accuracy of the optimised similarity model with respect to lower and upper bounds. It addresses the following questions:

- *Do GRC similarity models optimised on gold standard data offer an improvement over the baseline?*
- *How does GRC using the optimised similarity model compare to human performance on the task?*

This evaluates the contribution of the similarity model developed here with respect to a simple baseline approach on the ACE 2005 news test data. It also compares to a human upper bound derived from the ACE double annotation (see Section 5.2.2).

5.4.2.2 Results

Table 5.5 contains precision (P), recall (R) and f-score (F) results. Rows in the table correspond to the lower bound based on a random partition of the data (LB:Rand), the clustering approach using LDA dimensionality reduction and the WED feature set (Cl:LDA) and the upper bound based on human agreement (UB:Hum). The best score for each evaluation measure is in bold. Systems that are statistically distinguishable from the best for the given measure (i.e., $p \leq 0.05$) are underlined. The f-score results are mixed as to whether the Cl:LDA model outperforms the lower bound, with p values of 0.1094 for $F_{1:1}$ and 0.0078 for F_{pw}. Other results are consistent across the 1-to-1 and pairwise evaluation measures, however, with Cl:LDA being significantly better than LB:Rand in terms of recall and the two being statistically indistinguishable in terms of precision. Recall (especially R_{pw}) is relatively high with respect to precision for the LDA model. Again, this can be attributed to small values of hyperparameters (further discussion in Section 5.4.3 below).

There is room for improvement with respect to the human upper bound. Cl:LDA performs significantly worse in terms of all measures except pairwise recall (R_{pw}). It should probably be noted that inter-annotator agreement on ACE is a very strong upper bound for the GRC task as the annotators are given detailed guidelines of a specific relation type schema. The GRC task, on the other hand, is not guided by a pre-defined schema. Nevertheless, the comparison is informative to get a rough of idea of human performance on the task.

a) 1-to-1 F-Score ($F_{1:1}$)

	$P_{1:1}$	$R_{1:1}$	$F_{1:1}$
LB:Rand	0.414	0.429	0.485
Cl:LDA	**0.564**	**0.634**	**0.591**
UB:Hum	0.969	0.923	0.966

b) Pairwise F-Score (F_{pw})

	P_{pw}	R_{pw}	F_{pw}
LB:Rand	0.509	0.366	0.415
Cl:LDA	**0.523**	**0.875**	**0.646**
UB:Hum	0.946	0.937	0.941

Table 5.5: *Precision (P), recall (R) and f-score (F) results for LDA-reduced similarity model (Cl:LDA) with respect to lower (LB:Rand) and upper (UB:Hum) bounds on news test set. The best score in each column is in bold and those that are statistically distinguishable from the best are underlined.*

5.4.3 Experiment 3: GRC Across Domains

5.4.3.1 Method

The final experiment addresses the claim of modification-free domain adaptation (i.e., that models achieve comparable accuracy when transferred, without modification of model parameters, across domains). Specifically, it poses the following question:

- *Do GRC similarity models generalise across data sets and domains?*

Here, the performance of the various models are compared on the news and biomedical domains. These models are optimised on the news development data (ACE 2004) and applied directly to the news (ACE 2005) and biomedical (BioInfer) test sets without modification. Results for the lower bound are also presented for comparison.

5.4.3.2 Results

Table 5.6 contains precision (P), recall (R) and f-score (F) results. Rows correspond to the lower bound (LB:Rand), the unreduced clustering approach (Cl:None), the SVD-reduced approach (Cl:SVD) and the LDA-reduced approach (Cl:LDA). All clustering approaches here use the WED feature set. The best score for each evaluation measure is in bold and systems that are statistically distinguishable from the best (i.e., $p \leq 0.05$) are underlined. Table 5.6(a) contains results for the news domain development set (ACE 2004); Table 5.6(b) contains results for the news domain test set (ACE 2005); and Table 5.6(c) contains results for the biomedical domain test set (BioInfer).

In terms of the f-score results of the clustering systems, the LDA-reduced similarity model achieves the highest scores in most combinations of sub-domains and

a) ACE 2004 (News Development Set)

	$P_{1:1}$	$R_{1:1}$	$F_{1:1}$		P_{pw}	R_{pw}	F_{pw}
LB:Rand	0.583	0.357	0.437	LB:Rand	0.521	0.295	0.372
Cl:None	0.720	0.511	0.591	Cl:None	**0.616**	0.414	0.486
Cl:SVD	**0.726**	0.540	0.609	Cl:SVD	0.593	0.472	0.505
Cl:LDA	0.692	**0.685**	**0.683**	Cl:LDA	0.551	**0.923**	**0.676**

b) ACE 2005 (News Test Set)

	$P_{1:1}$	$R_{1:1}$	$F_{1:1}$		P_{pw}	R_{pw}	F_{pw}
LB:Rand	0.414	0.429	0.485	LB:Rand	0.509	0.366	0.415
Cl:None	**0.674**	0.566	**0.607**	Cl:None	**0.552**	0.511	0.513
Cl:SVD	0.663	0.555	0.599	Cl:SVD	0.543	0.523	0.518
Cl:LDA	0.564	**0.634**	0.591	Cl:LDA	0.523	**0.875**	**0.646**

c) BioInfer (Biomedical Test Set)

	$P_{1:1}$	$R_{1:1}$	$F_{1:1}$		P_{pw}	R_{pw}	F_{pw}
LB:Rand	0.655	0.444	0.525	LB:Rand	0.597	0.374	0.455
Cl:None	0.729	0.522	0.600	Cl:None	**0.644**	0.457	0.526
Cl:SVD	**0.765**	0.596	0.663	Cl:SVD	0.639	0.586	0.587
Cl:LDA	0.720	**0.705**	**0.708**	Cl:LDA	0.606	**0.779**	**0.672**

Table 5.6: *Comparison of precision (P), recall (R) and f-score (F) results on news and biomedical test sets. Rows correspond to the lower bound (LB:Rand), unreduced (Cl:None), SVD-reduced (Cl:SVD) and LDA-reduced (Cl:LDA) models. The best score in each column is in bold and those that are statistically distinguishable from the best are underlined.*

evaluation measures. And it is significantly better than the baseline across all combinations. The LDA-reduced model is significantly better than the unreduced and SVD reduced models in terms of F_{pw} on both the news development and test sets, though not on the biomedical test set. Taking a perfect upper bound (i.e., f-scores of 1), the LDA-reduced system achieves error rate reductions with respect to the SVD-reduced system of 34.5%, 26.6% and 20.6% respectively on the ACE 2004, ACE 2005 and BioInfer data sets. In terms of recall, however, the LDA-reduced model is significantly better than the unreduced model for all combinations except in terms of $R_{1:1}$ on the news test set. Again, the effect of the hyperparameters can be observed in the relatively high recall for the LDA-reduced model. Here, the small values of α (means across sub-domains of 0.63, 0.67 and 0.65 respectively for the ACE 2004, ACE 2005 and BioInfer data sets) can be expected to result in skewed topic distributions, which subsequently lead to skewed distributions over clusters. This effect can be observed in terms of the very strong negative correlation between values of α and pairwise recall (Pearson's r of -0.686, -0.733 and -0.865 respectively for the ACE 2004, ACE 2005 and BioInfer data sets).

While Cl:LDA is significantly better than Cl:SVD on both news data sets in terms of F_{pw}, the fact that the performance of the SVD and LDA systems is similar suggests that a choice between them can be made freely based on other criteria. SVD-reduced similarity models may be preferable in some cases, e.g. where scalability rules out LDA (Turney, 2006). However, based on the results here, LDA performs at least as well as SVD-reduced models and arguably better (in terms of F_{pw}). Therefore, because of the interpretability argument above (Section 5.3.2.3), the LDA-reduced similarity model is preferred for the extrinsic evaluation in Chapter 6.

5.5 Analysis

5.5.1 Characterisation of Entity Pair Sub-Domains and Performance

Table 5.7 contains, for each entity pair sub-domain, the sub-domain size (N), the type-to-token ratio (TTR), the number of relation mentions ($|\mathcal{K}|$) and the entropy of the relation type distribution ($H(K)$). Type-to-token ratio (TTR) is the number of features divided by the number of feature instances and indicates how much repetition there is in features. Since TTR can vary depending on the denominator (i.e., the number of tokens), it is computed on a random sample of 30 features from each sub-domain.

a) ACE 2005 (News Test Set)

	Sub-Domain Statistics				Sub-Domain $F_{1:1}$			Sub-Domain F_{pw}		
SD	N	TTR	$\|\mathcal{K}\|$	$H(K)$	None	SVD	LDA	None	SVD	LDA
F-G	41	0.865	3	1.520	0.718	0.756	0.701	0.522	0.540	0.565
F-P	41	0.912	2	0.872	0.707	0.684	0.536	0.631	0.696	0.650
G-G	109	0.862	2	0.411	0.711	0.696	0.864	0.674	0.656	0.890
G-O	98	0.856	2	0.999	0.608	0.583	0.352	0.570	0.532	0.658
G-P	322	0.905	3	1.345	0.631	0.581	0.537	0.614	0.452	0.596
O-P	195	0.912	7	1.685	0.360	0.420	0.665	0.227	0.321	0.630
P-P	71	0.877	4	1.524	0.513	0.474	0.479	0.354	0.430	0.531
μ	125	0.884	3.3	1.194	0.607	0.599	0.591	0.513	0.518	0.646

b) BioInfer (Biomedical Test Set)

	Sub-Domain Statistics				Sub-Domain $F_{1:1}$			Sub-Domain F_{pw}		
SD	N	TTR	$\|\mathcal{K}\|$	$H(K)$	None	SVD	LDA	None	SVD	LDA
A-N	42	0.863	3	1.198	0.657	0.658	0.656	0.442	0.480	0.486
N-N	33	0.843	4	1.558	0.726	0.765	0.583	0.613	0.653	0.423
P-C	130	0.855	2	0.737	0.585	0.755	0.673	0.634	0.640	0.751
P-F	187	0.769	2	0.364	0.671	0.667	0.819	0.634	0.633	0.791
P-P	694	0.898	3	1.355	0.508	0.638	0.670	0.409	0.675	0.666
P-S	126	0.813	3	0.940	0.495	0.626	0.848	0.445	0.515	0.817
R-B	89	0.856	3	0.931	0.555	0.534	0.708	0.505	0.512	0.771
μ	186	0.842	2.9	1.012	0.600	0.663	0.708	0.526	0.587	0.672

Table 5.7: *Breakdown of f-scores by sub-domain with number of relation mentions (N), type-to-token ratio (TTR), number of gold standard relation types ($|\mathcal{K}|$), and entropy of relation type distribution ($H(K)$).*

This is repeated ten times and averaged. Note also that TTR is computed on the same features that are used for clustering, i.e. the list of features for TTR is simply the concatenated list of features (with repetition) for each relation mention within a sub-domain. Entropy is represented as $H(K)$, where K is a random variable encoding the gold standard class distribution. $H(K)$ can be interpreted as a measure of the uniformity of a distribution. Low $H(K)$ indicates a more spiked distribution while high $H(K)$ indicates a more uniform distribution. Table 5.7 also contains the sub-domain f-score results for the unreduced (None), SVD-reduced (SVD) and LDA-reduced (LDA) systems with the WED feature set. Rows in the table correspond to the entity pair sub-domains. Entity types for the news data include FACILITY/VEHICLE/WEAPON (F), GEOGRAPHICAL/POLITICAL/LOCATION (G), ORGANISATION (O) and PERSON (P). Entity types for the biomedical data include AMINO-ACID (A), SUBSTANCE (B), PROTEIN-COMPLEX (C), PROTEIN-FAMILY (F), NUCLEIC-ACID (N), INDIVIDUAL-PROTEIN (P), SOURCE (R) and PROTEIN-SUBSTRUCTURE (S). Inspection of the table seems to suggest that the most difficult sub-domains (i.e., those with the lowest f-scores) have both high TTR (little repetition in features) and high $H(K)$ (relation type distribution close to uniform). This is exemplified by the ACE 2005 P-P and the BioInfer P-P sub-domains, where TTR and $H(K)$ are above average and most f-scores are below average. By contrast, sub-domains that have low TTR or low $H(K)$ tend to be easier.

These tendencies are summarised in the correlation analysis in Table 5.8, where columns correspond to the various sub-domain characteristics described above and rows correspond to the unreduced (None), SVD-reduced (SVD) and LDA-reduced (LDA) systems respectively. The values in the table correspond to the correlation (Pearson's r).[13] For example, the first data cell of the first table contains the correlation (-0.337) across sub-domains between the the sub-domain size (N) and the 1-to-1 f-score result for the unreduced system (None). Following conventions in the literature for effect size of Pearson's r (e.g., Cohen, 1988; Coolican, 2004), values over 0.10 (typeset in italicised bold font) are considered to indicate a small effect, values over 0.30 (typeset in bold font) are considered to indicate a medium effect and values over 0.50 (underlined bold font) a strong effect. Negative correlation values support the conclusion that sub-domains with high TTR (little feature repetition) and high $H(K)$ (relation type distribution close to uniform) correspond to low f-scores. In addition,

[13] A detailed correlation analysis using multiple regression, which accounts for covariance between the independent variables, would be interesting here. However, there are only seven sub-domains; not enough data points for reliable results (e.g., Coolican, 2004, p 464).

a) ACE 2005 (News Test Set)

$F_{1:1}$	N	TTR	$\lvert\mathcal{K}\rvert$	$H(K)$	F_{pw}	N	TTR	$\lvert\mathcal{K}\rvert$	$H(K)$
None	**-0.337**	**-0.378**	**-0.907**	**-0.633**	None	-0.064	***-0.279***	**-0.929**	**-0.786**
SVD	**-0.414**	**-0.375**	**-0.764**	**-0.544**	SVD	**-0.521**	***-0.265***	**-0.856**	**-0.835**
LDA	-0.019	-0.049	**-0.544**	***-0.269***	LDA	-0.035	***-0.273***	**-0.347**	**-0.861**

b) BioInfer (Biomedical Test Set)

$F_{1:1}$	N	TTR	$\lvert\mathcal{K}\rvert$	$H(K)$	F_{pw}	N	TTR	$\lvert\mathcal{K}\rvert$	$H(K)$
None	**-0.518**	**-0.315**	***0.192***	0.082	None	**-0.437**	**-0.520**	**-0.315**	**-0.424**
SVD	***-0.165***	-0.063	0.019	***0.166***	SVD	**0.525**	0.056	-0.099	0.068
LDA	-0.008	**-0.690**	**-0.517**	**-0.726**	LDA	***0.181***	**-0.387**	**-0.688**	**-0.770**

Table 5.8: *Spearman's r correlation analysis comparing f-scores of unreduced (None), SVD-reduced (SVD) and LDA-reduced (LDA) systems to number of relation mentions (N), type-to-token ratio (TTR), number of gold standard relation types ($\lvert\mathcal{K}\rvert$) and entropy of relation type distribution ($H(K)$).*

negative correlation values suggest that sub-domains with high $\lvert\mathcal{K}\rvert$ (many relation types) also correspond to low f-scores. This is consistent across sub-domains and measures for the unreduced and LDA-reduced systems though not consistent across sub-domains for the SVD-reduced system.

5.5.2 Error Analysis

This section contains an analysis that aims to characterise the types of errors made by the LDA-reduced system with the WED feature set. Like the pairwise accuracy measures (described in Chapter 3), these are based on pairs of data points, i.e. whether two data points have both the same system cluster and the same gold standard class. Specifically, the analysis here looks at false positive and false negative errors. False positive errors are pairs of data points that are in the same cluster but do not have the same gold standard class (i.e., are clustered together when they should not have been). False negative errors are pairs of data points that have the same gold standard class but are not in the same cluster (i.e., are not clustered together when they should have been).

5.5.2.1 False Positives

Inspection of false positive (FP) errors allows the characterisation of common causes of low precision. The focus is on the most frequent false positive errors, defined as those that account for more than 15% of the total number of clustered instance pairs in the system output (i.e., pairs of relation mentions that are part of the same system cluster) for any of the entity pair sub-domains. Percentages are calculated with respect to the total number of clustered instance pairs in the system output to quantify the impact on precision.

LOCATED relation types (describing the physical location of an entity) account for a very large number of wrongly clustered data points. These are frequently confused with several relation types. In the ACE 2005 G-P sub-domain, for instance, 36.4% of clustered instance pairs consist of LOCATED relation mentions wrongly paired with EMPLOYMENT relation mentions. Consider the following two sentence excerpts, the first containing a LOCATED relation mention between "geraldo rivera" and "iraq" and the second containing a EMPLOYMENT relation mention between "John Negroponte" and "US":

1. "[^{Person} geraldo rivera] may not be kicked out of [^{GPL} iraq]."

2. "[^{GPL} U.S.] Ambassador [^{Person} John Negroponte]"

The first is difficult because the fact that "geraldo rivera" is in "iraq" has to be inferred from a statement about whether he will be leaving. In the second, the EMPLOYMENT relation mention has to be inferred from the title of the job (i.e., ambassador). Other common FP errors involving LOCATED relation mentions include incorrect clustering with SUBSIDIARY and USER-OWNER-INVENTOR-MANUFACTURER relation mentions. Errors with SUBSIDIARY relation mentions account for 50% of clustered instance pairs in the ACE 2005 G-O sub-domain while errors with USER-OWNER-INVENTOR-MANUFACTURER relation mentions account for 44.6% of pairs in the ACE 2005 F-P sub-domain.[14]

FAMILY relation mentions are frequently confused with two other types of relations. In the ACE 2005 P-P sub-domain, for instance, 15.4% of clustered instance pairs

[14] GEOGRAPHICAL relation mentions (describing *in*, *at* or *part-of* relations) are similar to LOCATED relation mentions in that they are often wrongly clustered together with USER-OWNER-INVENTOR-MANUFACTURER relation mentions, accounting for 33.4% of clustered instance pairs in the ACE 2005 F-G sub-domain. GEOGRAPHICAL relation mentions are also wrongly clustered together with LOCATED relation mentions in the ACE 2005 G-G and ACE 2005 F-G sub-domains, accounting respectively for 15.7% and 14.8% of pairs.

consist of FAMILY relation mentions wrongly paired with LASTING-PERSONAL relation mentions (describing other long-term personal relations e.g. friendship). Consider the following two fragments, the first containing a FAMILY relation mention between "her" and "scott" and the second containing a LASTING-PERSONAL relation mention between "her" and "anna":

1. "[^{Person} her] husband [^{Person} scott]"

2. "[^{Person} her] friend [^{Person} anna]"

The difference between these two relation mentions is actually very subtle. In fact, for many applications of generic relation extraction, it is probably not detrimental to have these two relation mentions clustered together. FAMILY relation mentions are also frequently wrongly clustered with BUSINESS relation mentions in the ACE 2005 P-P sub-domain, accounting for 29.3% of clustered instance pairs.

EMPLOYMENT and MEMBERSHIP relation mentions are frequently wrongly clustered together in the ACE 2005 O-P sub-domain, accounting for 22.6% of clustered instance pairs. Consider the following sentence excerpts, the first containing an EMPLOYMENT relation mention between "Kofi Annan" and "U.N." and the second containing a MEMBERSHIP relation mention between "whitman" and "republican":

1. "[^{Organisation} U.N.] Secretary General [^{Person} Kofi Annan]"

2. "[^{Person} whitman] did consider herself sort of the [^{Organisation} republican] environmentalist"

Again, the difference is subtle and probably not essential to applications of generic relation extraction.

In the biomedical data, CAUSAL and IS-A relation mentions are frequently wrongly clustered together, accounting respectively for 34.3%, 20.5% and 15.1% of clustered instance pairs in the A-N, N-N and P-S sub-domains. CAUSAL and OBSERVATION relation mentions are also frequently confused, accounting for 24.5% and 31.9% of clustered instance pairs in the P-P and R-B sub-domains. In the P-C sub-domain, 34.4% of clustered instance pairs consist of wrongly clustered CAUSAL and PART-OF relation mentions. And, finally, in the N-N sub-domain, 31.1% of clustered instance pairs consist of wrongly clustered PART-OF and IS-A relation mentions.

5.5.2.2 False Negatives

Inspection of false negative (FN) errors allows the characterisation of common causes of low recall. The rest of this section focuses on the most frequent false negative (FN) errors, defined as those that account for more than 5% of the total number of grouped instance pairs in the gold standard annotation (i.e., pairs of of relation mentions that are part of the same gold standard class) for any of the entity pair sub-domains. Percentages are calculated with respect to the total number of grouped instance pairs in the gold standard annotation to quantify the impact on recall.

LOCATED relation types (describing the physical location of an entity) account for a large number of FN errors. In the ACE 2005 F-P and F-G sub-domains respectively, these account for 21.2% and 8.4% of instance pairs that have the same gold standard class. Consider the following two sentence excerpts, the first containing a LOCATED relation mention between "allan chernoff" and "new york stock exchange" and the second between "they" and "ramstein air base":

1. "[^{Person} allan chernoff] live from the [^{FVW} new york stock exchange]"

2. "[^{Person} they]'ll be arriving at [^{FVW} ramstein air base]"

Both of these are somewhat difficult due the fact that they are arguably event mentions that are meant to be interpreted as relation mentions (i.e., a *reporting* event mention in the first and an *arriving* event mention in the second). Furthermore, the second relation mention also shows up twice in the sample of false positive errors. Both times it is clustered with USER-OWNER-INVENTOR-MANUFACTURE relation mentions, which could also be considered a valid relation type.

GEOGRAPHICAL relation types (describing *in*, *at* or *part-of* relations) also account for a large number of FN errors. In the ACE 2005 F-G and G-G sub-domains respectively, these account for 12.6% and 8.4% of instance pairs that have the same gold standard class. Consider the following sentence excerpts, the first of which has a GEOGRAPHICAL relation mention between "Paris" and "European" and the second between "Lahaina" and "Hawaii":

1. "Washington's anger with [^{GPL} European] resistance to the campaign was focused more on [^{GPL} Paris]"

2. "[^{GPL} Lahaina], [^{GPL} Hawaii]"

Here, the first is rather difficult as it is not explicitly stated. Identifying this relation mention requires either 1) inferring that focusing anger with European resistance on Paris means that Paris is part of Europe or 2) world knowledge.

FAMILY relation FN errors are also common, accounting for 24.8% of instance pairs that have the same gold standard class in the ACE 2005 P-P sub-domain. Consider the following sentence excerpts, the first containing a FAMILY relation mention between "her" and "sara" and the second between "tariq aziz" and "ziad":

1. "she said she could no longer cope with [*Person* her] daughter daughter's learning, disability prps [*Person* sara]"

2. "[*Person* tariq aziz]'s sun [*Person* ziad]"

Both of these relation mentions are from the broadcast news data. In the first, the disfluencies in the transcription (i.e., the repetition of the word "daughter", the misplaced comma, the non-word token "prps" and the lack of capitalisation) lead to errors in the dependency path. In the second, by contrast, the disfluencies in the transcription (i.e., the confusion of the words "sun" and "son") do not lead to errors in the dependency path. The coincidence of the FN error and the bad dependency path is highly representative as 80% of the sampled FN errors in the ACE 2005 data include at least one entity mention pair with a bad or noisy dependency path.[15]

In the biomedical data, IS-A relation mentions account for the largest percentage of FN errors. In the A-N and N-N domains respectively, these account for 51.4% and 23.3% of instance pairs that have the same gold standard class. Consider the following two sentence excerpts, the first containing a IS-A relation mention between "member of the cofilin/ADF family" and "twinstar locus" and the second between "subunits of which" and "viral UL5, UL8 and UL52 genes":

1. "A similar phenotype was seen in testes treated with cytochalasin B and has been noted previously in mutants at the [*NucleicAcid* twinstar locus], a gene that encodes a Drosophila [*AminoAcid* member of the cofilin/ADF family]"

2. "Herpes simplex virus type I expresses a heterotrimeric helicase-primase, the [*AminoAcid* subunits of which] are encoded by the [*NucleicAcid* viral UL5, UL8 and UL52 genes]"

[15] It is not evident that the parser does worse on broadcast news. This can be summarised by calculating the correlation between a variable coding whether the dependency path was good or bad and a second variable coding whether the relation mention came from broadcast news or newswire data. The resulting phi correlation value for the ACE 2005 data is 0.032.

In the first, a bad dependency path leads to many extraneous features. In the second, the dependency path is correct, consisting of an *object* relation from the word "encoded" to the word "subunits" and a *by-subject* relation from the word "encoded" to the word "genes". Again, this coincidence of the FN error and the bad dependency path is representative as 56% of the sampled FN errors in the BioInfer data included at least one entity mention pair with a bad or noisy dependency path. The fact that this is much lower than the 80% reported for the news data is due at least in part to the fact that a large proportion of entity mention pairs in the BioInfer data are nominal modifier relations (e.g., the PART-OF relation mention between "histone" and "H3" in "[^{ProteinFamily} histone] [^{Protein} H3]").

PART-OF and CAUSAL relation mentions also account for many FN errors in the BioInfer data. In the P-F, N-N and P-C sub-domains, PART-OF FN errors account respectively for 26.2%, 14.0% and 12.0% of instance pairs that have the same gold standard class. And in the N-N and R-B sub-domains, CAUSAL FN errors account for 11.6% and 10.7% of instance pairs that have the same gold standard class.

5.6 Summary and Future Work

This chapter presented experiments addressing the generic relation characterisation task (GRC), comparing similarity models for clustering entity mention pairs by relation type. A novel feature set was introduced for the task based on syntactic features from governor-dependency parses and two dimensionality reduction techniques were compared. The first dimensionality technique was singular value decomposition (SVD), a linear algebraic method that has proved successful in the language processing and information retrieval literature. The second version was latent Dirichlet allocation (LDA), a probabilistic generative analogue of SVD. The dimensionality reduction approaches were compared to a similarity model with no dimensionality reduction and to a baseline that creates a random partition.

Experiments suggest that the LDA-reduced model successfully incorporates a larger and more interdependent feature set than the unreduced and SVD-reduced models. This was explained in terms of the LDA hyperparameters, which control the impact of sparsity. Across domains, the LDA-reduced model is significantly better than the SVD-reduced model in terms of pairwise f-score on both the ACE 2004 and ACE 2005 data, obtaining reductions in the error rate with respect to perfect performance (i.e., f-scores of 1) of 34.5% and 26.6% respectively. The error rate reduction is also high

on the BioInfer data (20.6%), though the the difference was not found to be significant. The LDA-reduced system does especially well in terms of recall. Again, this was attributed to small values for the LDA hyperparameters leading to skewed topic distributions, which subsequently lead to skewed distributions over clusters.

A characterisation of entity pair sub-domains suggested that the clustering systems struggle most on tasks with little repetition in features, relation type distributions close to uniform and/or many relation types. Error analysis identified commonly confused relation types. False positive errors tended to include relation mentions that required inference or used figurative language. In addition, some false positive errors, while incorrect according to the gold standard, were deemed to constitute subtle differences between relation types that are probably not essential to applications of generic relation extraction. Among false negative errors, bad dependency paths were highly prevalent. Other false negative errors were due to transcription errors in the broadcast news data and to difficult relation mentions that required inference or used figurative language.

While the LDA-reduced model is significantly better than SVD-reduced model on both news data sets in terms of F_{pw}, the fact that the performance of the SVD and LDA systems is similar suggests that a choice between them can be made freely based on other criteria. SVD-reduced similarity models may be preferable in some cases, e.g. where efficiency is more important than accuracy. However, based on the results here, LDA performs at least as well as SVD-reduced models and arguably better (in terms of F_{pw}). Therefore, because of the interpretability argument above (Section 5.3.2.3), the LDA-reduced similarity model is used for the extrinsic evaluation in Chapter 6.

The experiments with dependency path features here show that this is a useful source of information for GRC. However, the feature set makes limited use of the information in dependency parses. Representations could be extended, for example, to incorporate dependency triples consisting of two word tokens and the relation between them or a full dependency path representation like that used by Lin and Pantel (2001) for discovery of inference rules for question answering. It may also be useful to incorporate information from outside the dependency path. This could could help capture relation type information, for instance, in cases where relation mentions exist between conjoined entity mentions (e.g., in the sentence ““Teammate [^{Person} Kobe Bryant] said he would not be surprised if [^{Person} O'Neal] had a big game Saturday”, where the word “teammate” describes the relation mention but is not on the dependency path).

There are also many options for extending models due to the flexible nature of probabilistic topic modelling. Some possibilities include: 1) non-parametric models

where the number of topics is sampled (e.g., Blei et al., 2004; Teh et al., 2004), 2) variations of the model topology that model topics with respect to entity mention pairs akin to author-topic and author-recipient-topic models (e.g., Rosen-Zvi et al., 2004; McCallum et al., 2004), and 3) approaches that integrate coreference into the generic relation extraction task. Coreference information could be integrated by augmenting the feature space based on the output of a preceding coreference module akin to related work in summarisation (e.g., Steinberger et al., 2005). Coreference information could also be integrated using joint or iterative models that use coreference information to inform relation extraction models and vice versa akin to related work in named entity recognition (e.g., Wellner et al., 2004). With respect to the latter approach, one could use distributional information over relation types and related entity mentions to contribute to similarity models for entity coreference.

The next chapter will demonstrate that the generic relation extraction approach developed in this thesis is useful for an extrinsic summarisation task. However, as the error analysis demonstrated, the output is somewhat noisy. Another possible use of the generic approaches developed here is as a way of initialising a fully bottom-up active learning approach. This could be easily achieved by using human annotators to introduce pairwise constraints, which can be incorporated using semi-supervised clustering approaches based on learnt similarity measures (e.g., Klein et al., 2002; Xing et al., 2003; Bilenko et al., 2004) or by assigning arbitrarily high similarity values to pairs annotated as having the same type and arbitrarily low values to pairs annotated as not having the same type (e.g., Blum and Chawla, 2001).

Chapter 6

Generic Relation Extraction and Multi-Document Summarisation

Generic relation extraction is not necessarily an end in itself but can be used to enhance applications such as automatic summarisation. To this end, experiments are reported on an extractive multi-document summarisation task, using representations based on generic relation extraction. This serves as an extrinsic evaluation of end-to-end GRE based on the models developed in this thesis, showing significantly improved performance over a non-trivial baseline based on *tf*idf*-weighted words. Furthermore, the experiments here demonstrate that models tuned on relation extraction data achieve comparable relative accuracy when used for summarisation.

6.1 Introduction

The goal of summarisation is to take an information source, extract content from it, and present the most important content in a condensed form (Mani, 2001). Summaries can be intended to convey all of the important content from the source or can be intended just to help the reader decide whether to look at the original (Borko and Bernier, 1975). Abstracts of academic publications are an example of (generally indicative) summaries, which are often written by the authors (cf. e.g., ANSI: American National Standards Institute, 1997). In some professional fields, summaries are often prepared by information management professionals at publishing or archiving organisations. The Incorporated Council for Law Reporting[1], for example, publishes human-written

[1] `http://www.lawreports.co.uk/`

summaries of UK court proceedings, which are important due to the central role of precedent in English common law.

The field of automatic summarisation (cf., Endres-Niggemeyer, 1998; Mani, 2001; Spärck Jones, 2007) aims to create tools that address various summarisation tasks with minimal human intervention. Mani (2001, pp. 19–20) lists a number of applications from the literature, including: news summarisation from multiple online sources (e.g., Columbia Newsblaster[2], NewsInEssence[3]); assistants for patient-focused access to medical literature (e.g., McKeown et al., 1998; Elhadad and McKeown, 2001); meeting summarisation (e.g., Waibel et al., 1998; Murray et al., 2005a); and re-presentation for devices with small screens (e.g., Nakao, 2000; Corston-Oliver, 2001).

Approaches to summarisation are frequently divided into two main camps: abstractive and extractive. Abstractive approaches create conceptual representations of the source document based on deep linguistic pre-processing and thus require summary generation techniques that create novel output (i.e., text strings that are not necessarily found in the input). Extractive approaches create representations of the source document that are generally based on an easily identified text sub-unit such as sentences or paragraphs. These representations are then used to identify representative or important snippets of text to place in the summary. Research on extractive approaches constitutes a very large majority of current work on automatic summarisation. This is due to the fact that extractive systems are more general in that they do not require much in the way of domain-specific resources for interpretation, which also makes them relatively easy to develop. For the purposes of the current work, extractive summarisation provides a convenient framework for comparing models of text content.

Another important distinction is between extractive approaches that use unsupervised salience functions and those that learn salience functions from annotated training material. The former generally use functions based on frequency counts of extract indicator features. Specific examples include the Luhn (1958), Edmundson (1968) and Filatova and Hatzivassiloglou (2004) systems, which are described in Section 6.2 below. The latter generally use supervised machine learning algorithms to learn a function that maps features of a text unit to a salience prediction. Specific examples include the systems described by Kupiec et al. (1995) and Teufel and Moens (1997), which learn summary classifiers using naïve Bayes to incorporate various features including cue phrase, location, sentence length and term importance information. Supervised

[2] http://newsblaster.cs.columbia.edu/

[3] http://lada.si.umich.edu:8080/clair/nie1/nie.cgi

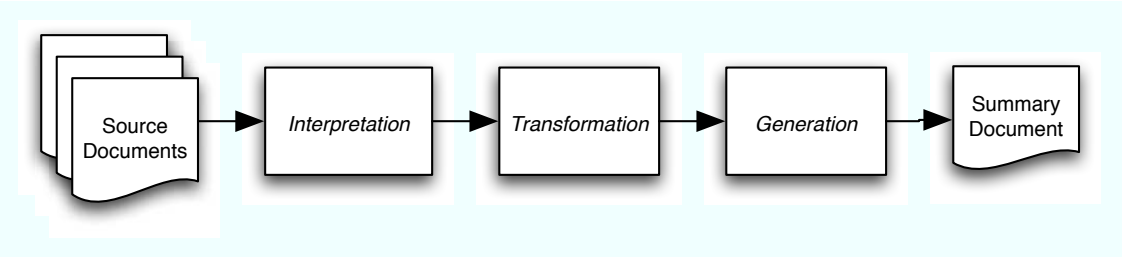


Figure 6.1: *Main sub-tasks of automatic summarisation.*

approaches have the advantage of being able to incorporate diverse information into a single salience function. However, they require training material where sentences are annotated for extract-worthiness.

Following Spärck Jones (1999, 2007), summarisation systems can be characterised with respect to their approach to the three main sub-tasks in Figure 6.1. The input consists of the source document (or a collection of source documents in the case of multi-document summarisation). The first step (interpretation) creates a representation of the source document by performing some level of interpretation. A simple approach here would represent sentences by their tokens (i.e., as an unordered bag-of-words). The next step (Transformation) is the compaction step where the source representation is converted into the summary representation, e.g. by identifying sentences whose word subsets are most representative of the full text. Finally, in the generation step, the output summary is created. In the case of sentence extraction, this involves preparation which includes various operations to maximise coherence such as ensuring that entity references are comprehensible and arranging the sentences in a sensible order.

The current work uses a sentence extraction framework to evaluate relation-based representations based on GRE output and will refer to summarisation via sentence extraction using unsupervised salience functions by default. From a summarisation perspective, the motivation is to explore GRE as a novel knowledge stream and explore the type of summaries where it will be beneficial. Section 6.2 situates the current work with respect to related source representations from the summarisation literature. Section 6.3 describes the setup for the experimental evaluation. The models compared here are described in Section 6.4. The experiments are present in Sections 6.5.1, 6.5.2 and 6.5.3. And the analysis of results is presented in Sections 6.6.

6.2 Review

The interpretation and representation of source documents is an important aspect of automatic extractive summarisation which has experienced a slow but steady evolution in the literature. In seminal work, Luhn (1958) introduces a representation based on content words. These are defined as non-function words from the source document that are neither too frequent nor too infrequent. Luhn uses frequency to weight content words and extracts sentences with the highest combined content scores to form the summary. Subsequent work adapted the *tf*idf* weighting scheme, where term frequency (*tf*) is combined with inverse document frequency (*idf*), an inverse measure of term occurrence across documents that serves to down-weight common words (Spärck Jones, 1972). Another representation used in early work is based on the position of sentences within a document. These features are based on the observation that important information tends to occur at the edges of e.g. documents and paragraphs (Baxendale, 1958; Brandow et al., 1995). In modern work, position- and *tf*idf*-based representations are often used as simple but non-trivial baselines, e.g. the DUC shared-tasks[4] and Filatova and Hatzivassiloglou (2004).

Other work has incorporated various cue phrase information into representations of the source documents (Edmundson, 1968; Rush et al., 1971). Edmundson, for example, use lists of words derived from an annotated training corpus that indicate whether a sentence salience score should be increased (bonus words) or decreased (stigma words). According to Edmundson's analysis, bonus lists tend to contain "cue words that are comparatives, superlatives, adverbs of conclusion, value terms, relative interrogatives and causality terms" while stigma lists contain "anaphoric expressions, belittling expressions, insignificant-detail expressions and hedging expressions". Later work explored the use of both single and multiple word indicators, which include bonus phrases like 'the purpose of this research is' and 'our investigation has shown that' as well as bonus words and stigma words.

In recent literature, representations have tended toward more abstract approaches based on linguistic structure, therefore moving toward underlying content rather than relying on surface structure. Barzilay and Elhadad (1997) and Lin and Hovy (2000), for example, present systems that generalise over surface structures by mapping words to more abstract conceptual representations based on synonymy. These can be built using lexical knowledge sources like WordNet (e.g., Barzilay and Elhadad, 1997) or

[4] http://www-nlpir.nist.gov/projects/duc/index.html

derived automatically from large corpora (e.g., Lin and Hovy, 2000). Other approaches use models of the underlying discourse structure to determine extract-worthiness. Mike et al. (1994) and Marcu (1997), for example, incorporate explicit models of the rhetorical relations between sentences in the text. A full rhetorical analysis of a text produces a tree structure where the the most important information is closer to the root of the tree. Marcu's algorithm uses this property to rank the extract-worthiness of text snippets (i.e., phrases that correspond roughly to the syntactic notion of clauses). In a related approach to discourse-based summarisation, Boguraev and Kennedy (1997) incorporate salience of referential noun phrases (e.g., 'he', 'priest') into their extraction function.

Yet other approaches tending toward more abstract representations of underlying content have tried to capture semantics through logical forms or templates. Tucker and Spärck Jones (2005), for instance, describe a system that is based on logical forms, derived from a parser which the authors use to extract predicate-argument structures. For example, interpretation of the following sentence:

> "Japanese investment in Asia is propelling the region toward economic integration."

produces the following predications:

`1:propel(B,D)`	`4:name_of(C,Asia)`	`7:integration(E)`
`2:investment(B)`	`5:in(B,C)`	`8:economic(E)`
`3:japanese(B)`	`6:region(D)`	`9:toward(D,E)`

The authors use these to build graphs linked by arguments (e.g., predication node `1` above is linked to nodes `2`, `3` and `5` by virtue of the fact that they all have `B` as an argument and is also linked to nodes `6` and `9` by virtue of the fact that they have `D` as a common argument). Links are also formed when predicates are the same or similar. When extracting sentences, the nodes with highest linkage according to several measures are selected first and the corresponding sentences extracted.[5]

This kind of logical approach is very interesting because it helps to incorporate underlying linguistic structure. However, it still relies on matching arguments and

[5] In addition to their sentence extraction approach, Tucker and Spärck Jones (2005) describe a phrase generation approach that lies somewhere on the continuum between purely extractive and purely abstractive systems. A number of other authors (e.g., McKeown and Radev, 1995; Barzilay et al., 1999; Elhadad and McKeown, 2001; Nobata et al., 2002; Vanderwende et al., 2004; Leskovec et al., 2005) also present work using representations derived from syntactic and semantic interpretation that move away from extractive summarisation towards approaches that are more abstractive in nature. These are not presented here due to the focus on extractive approaches as a test bed for comparing source representations based on information extraction.

predicates and thus does not do enough to abstract away from surface forms. As a remedy, the use of representations based on information extraction (IE) has been suggested. This is based on the notion that IE definitions of types for entities, relations and events provide a level of abstraction that may be more appropriate for automatic summarisation. McKeown et al. (1998), White and Cardie (2002) and Harabagiu and Maiorano (2002), for example, explore the use of IE-based representations for extractive summarisation: McKeown et al. incorporate patient characteristic templates for matching potential treatments to specific patients in a medical summarisation system; White and Cardie incorporate a bootstrapped IE system based on Autoslog (Riloff, 1996) for filling event templates; and Harabagiu and Maiorano incorporate a hybrid approach that uses conventional IE techniques for known topics and a more general approach based on WordNet for unknown topics.[6]

The problem with these systems is that they all use supervised approaches to IE that require that the IE templates be known in advance and additionally require significant investment in writing extraction rules or in annotating data for training. Where more general techniques are used, they still require domain-specific resources, e.g. White and Cardie's bootstrapping approach still requires that the extraction templates be known in advance and Harabagiu and Maiorano's approach depends on the WordNet lexical database, for which coverage is not guaranteed for arbitrary domains. Filatova and Hatzivassiloglou (2004) go a step further, introducing methods using more general IE representations that are not based on supervised learning. Given a named entity recogniser, the representation is automatically derived and consists of $<Ent, Connector, Ent>$ event triples, where connectors are verbs or action nouns (i.e., nouns that are hyponyms of event or activity in WordNet) that occur in between the two NEs. Thus, the approach aims to perform a simple generic IE task that the authors refer to as atomic events. This representation is shown to outperform a *tf*idf* baseline on the DUC 2001 data. As we will see in Section 6.4.3 below, Filatova and Hatzivassiloglou's approach has three main shortcomings.

The most obvious problem is the exclusive focus on simple atomic events (i.e., entity mention pairs with an intervening verbal connector), meaning that it will not be able to address tasks like biographical summarisation where relations are at least as important as events. Second, it suffers from the same problem as the Tucker and Spärck Jones approach in that it relies heavily on representations based on surface

[6]Comparable work using IE in the context of abstractive as opposed to extractive summarisation includes work by DeJong (1982), Hahn and Reimer (1999), White et al. (2001) and Saggion and Lapalme (2002).

forms, which are not capable of capturing latent semantic similarities. Finally, like the Harabagiu and Maiorano approach (though to a lesser extent), its performance is subject to the coverage of WordNet. In the rest of this chapter, a set of experiments are reported that compare the contribution of several possible knowledge streams to effective interpretation and representation of source documents for extractive summarisation. The experiments focus on IE-based representations including including the Filatova and Hatzivassiloglou event representation and a representation based on the GRE models developed in the previous chapters of this thesis.

6.3 The Task: Experimental Setup

For the sake of comparison, the experimental setup uses the same extraction algorithm, data and evaluation as Filatova and Hatzivassiloglou (2004). These are described in the remainder of this section.

6.3.1 Sentence Extraction as Set Cover

Given a source representation and weighting scheme, summarisation via sentence extraction generally proceeds by selecting the sentences with the highest weights and placing them in the summary until the summary meets the desired length (e.g., Luhn, 1958; Edmundson, 1968). However, this does not necessarily account for the possibility of redundancy (i.e., conceptual overlap between the selected sentences). To address this problem, Carbonell and Goldstein (1998) introduce the general purpose maximal marginal relevance (MMR) algorithm which combines an arbitrary salience function (e.g., based on unsupervised or supervised approaches to salience) with a second function that measures redundancy via overlap with already extracted text. Other recent work on algorithms has focused on salience, developing supervised learning approaches in which extractors are trained for given domains (e.g., Daumé III and Marcu, 2005; Daumé III, 2006) or approaches based on modern graph algorithms (e.g., Erkan and Radev, 2004; Yoshioka and Haraguchi, 2004; Mihalcea, 2005; Li et al., 2006).

By contrast, several recent papers have investigated various techniques that account for salience and redundancy using approximation algorithms for global inference from the literature. Filatova and Hatzivassiloglou (2004) introduce an approach based on algorithms for the set cover problem and compare several techniques for dealing with both salience and redundancy on the DUC 2001 multi-document summarisation

	c_1	c_2	c_3	c_4	c_5
t_1	1	1	0	1	1
t_2	1	0	0	1	0
t_3	0	1	0	0	1
t_4	1	0	1	1	1

Table 6.1: *Text $\times$ concept matrix for set cover approach to automatic summarisation (Filatova and Hatzivassiloglou, 2004).*

task. McDonald (2007) compares approaches based on a greedy search procedure similar to MMR, a dynamic programming solution (based on the knapsack problem) and integer linear programming (ILP). McDonald's results suggest that the dynamic programming solution performs as well as ILP in terms of accuracy on the DUC 2005 query-focused summarisation task and incurs a fraction of the computational resources required by ILP. As noted previously, the current evaluation adopts the Filatova and Hatzivassiloglou framework for the sake of comparing their atomic event models to the GRE-based models introduced here.

Filatova and Hatzivassiloglou (2004) define a general extraction model based on a mapping between *textual* units and *concepts*. To illustrate, consider the matrix in Table 6.1 where rows represent textual units (e.g., sentences, paragraphs) and columns represent concepts (e.g., words, events, relations) in the input text. Each concept is either absent or present in a given textual unit. Additionally, each concept has a weight associated with it. Looking at the problem in this way makes it natural to formulate it as follows: *the summary should select textual units such that there is maximal coverage of the salient conceptual units*.[7] This is essentially the maximum coverage problem, which has been shown to be reducible to the set covering problem and therefore to be NP-hard (e.g., Cormen et al., 2001). Approximation algorithms for set covering run in polynomial time or better (Hochbaum, 1997; Bienstock and Iyengar, 2004).

Filatova and Hatzivassiloglou define three greedy algorithms for extractive summarisation: a simple greedy algorithm and two versions of the greedy algorithm inspired by approximate solutions to the set covering problem. These can be parametrised

[7]While not considered in the current experiments, a more discourse-oriented approach could be derived within the set cover framework by down-weighting conceptual units that occur e.g. in portions of the source documents that describe background information, where text segments containing background information could be identified using a sentence-level rhetorical status classifier like that developed by Teufel and Moens (2002).

SUMMARISE : D,EXTRACT,UPDATE,k

1 $\mathcal{S} \leftarrow \{\}$
2 **while** $\sum_{t_i \in \mathcal{S}} \text{LENGTH}(t_i) < k$
3 $\quad t_j \leftarrow \text{EXTRACT}(D)$
4 $\quad \mathcal{S} \leftarrow \mathcal{S} \cup t_j$
5 $\quad D \leftarrow \text{UPDATE}(D, t_j)$
6 **return** $\mathcal{S}$

EXTRACT$_{default}$: D

1 $t_i \leftarrow \arg\max_{t_i \in rows(D)} \text{SCORE}(D, t_i)$
2 **return** t_i

UPDATE$_{static}$: D, t_i

1 $D \leftarrow \text{DELETE}(D, t_i)$
2 **return** D

Figure 6.2: *Generalised version of Filatova and Hatzivassiloglou (2004) function for extractive summarisation with default scoring function and static update function.*

in terms of the general SUMMARISE function in Figure 6.2. In addition to the text $\times$ concept matrix D and the maximum summary length k, this function takes an extraction function EXTRACT and a matrix update function UPDATE as input. The extraction function determines which text unit to select next for addition to the summary and makes use of a SCORE function. Filatova and Hatzivassiloglou simply calculate the sum over the concept weights for the given text unit, i.e.:

$$\text{SCORE} : D, t_i \mapsto \textbf{return} \sum_{c_j \in cols(D)} D[t_i, c_j] \tag{6.1}$$

The matrix update function determines how the text $\times$ concept matrix should be updated after a text unit is extracted. The function first initialises the summary $\mathcal{S}$ to the empty set. Then it enters a loop that continues until the summary reaches the desired length. Within the loop, a text unit is extracted and added to the summary after which the text $\times$ concept matrix is updated (i.e., to reflect the extracted text unit and, optionally, the extracted content). The output of the algorithm is a set $\mathcal{S}$ comprising the text units that make up the summary.

The first algorithm uses a simple greedy approach and is referred to as the static greedy algorithm (AL1). For this algorithm, the SUMMARISE function is invoked with the default extraction function and the static update function as defined in Figure 6.2. EXTRACT$_{default}$ selects the textual unit with the highest score and UPDATE$_{static}$ simply removes the row representing the extracted text unit from the text $\times$ concept matrix D. For the static greedy algorithm, the actual implementation of the extract function is made more efficient by calculating the text unit scores once and placing them in a heap data structure, then simply extracting text units from the top of the heap until the

```
UPDATE_adaptive : D, t_i
1  for each c_j ∈ cols(D)
2        if D[t_i, c_j] > 0
3              for each t_k ∈ rows(D)
4                    D[t_k, c_j] ← 0
5  D ← DELETE(D, t_i)
6  return D
```

Figure 6.3: *Matrix update function for adaptive greedy algorithm.*

summary reaches the desired length. The static greedy algorithm does not explicitly address redundancy in the summary.

The two adaptive versions of the greedy algorithm are referred to as the adaptive greedy algorithm (AL2) and the modified adaptive greedy algorithm (AL3). Intuitively, these aim to minimise redundancy in the summary by globally maximising the number of conceptual units covered in the output. For the AL2, the SUMMARISE function is invoked with the default extraction function from Figure 6.2 and the adaptive matrix update function ($\text{UPDATE}_{adaptive}$) defined in Figure 6.3. In addition to removing the row representing the extracted text unit from the text $\times$ concept matrix D, $\text{UPDATE}_{adaptive}$ iterates through the remaining text units and assigns zero weights to all concepts that are covered by the extracted text unit.

For AL3, the SUMMARISE function is invoked with the adaptive matrix update function from Figure 6.3 and the modified extraction function defined in Figure 6.4. $\text{EXTRACT}_{modified}$ first identifies the concept c_j not yet covered in the summary that has the highest overall weight in the text $\times$ concept matrix D. Then it selects the text unit t_k with the highest score from among the text units that contain concept c_j. This implementation of the modified extraction function is made more efficient by calculating the concept scores once and placing them in a heap data structure. Then, Line 1 of the modified extraction function is implemented as a loop that takes items from the top of the heap until a concept is found that is not yet covered in the summary.

For the experiments reported here, the text units (t) are sentences, the length function ($\text{LENGTH}(t_i)$) is the number of word tokens in sentence t_i and the score function ($\text{SCORE}(D, t_i)$) is the sum of concept weights defined in Equation 6.1. What remains is to define the mapping from text to conceptual units. This corresponds to the interpretation step in Figure 6.1, which takes the raw document collection as input and

$\text{EXTRACT}_{modified} : D$

1 $c_j \leftarrow \arg\max_{c_j \in cols(D)} \sum_{t_i \in rows(D)} D[t_i, c_j]$

2 $t_k \leftarrow \arg\max_{t_k \in rows(D) \& D[t_k, c_j] > 0} \text{SCORE}(D, t_k)$

3 **return** t_k

Figure 6.4: *Extraction function for modified greedy algorithm.*

outputs the conceptual representation. The various representations compared here are described below in Section 6.4. The code used is my own implementation, which achieves similar results to those reported by Filatova and Hatzivassiloglou (2004).

6.3.2 Data

The evaluation uses data created by the American National Institute for Standards and Technology[8] in the context of the document understanding conferences (DUC). DUC was a series of shared community evaluations that ran from 2001 through 2007 and explored various summarisation tasks including single document summarisation (e.g., DUC 2001, 2002), multi-document summarisation (e.g., DUC 2001, 2002, 2005), headline generation (e.g., DUC 2003, 2004) and query-focused summarisation (e.g., DUC 2005, 2006). The experiments here use the multi-document summarisation data from the DUC 2001 multi-document summarisation task (Harman and Marcu, 2001), which is the same data used by Filatova and Hatzivassiloglou (2004). This comprises 30 test document sets, which are made up of approximately 10 news stories. Each document set is collected by a human and focuses on a particular topic, though they vary in terms of coherence across documents. Example topics include the nomination of Clarence Thomas to the American Supreme Court, Neil Bush's role in the collapse of Silverado Savings and Loan and the Exxon Valdez oil spill. Gold standard summaries are provided for each document set for summary lengths of 50, 100, 200 and 400 words. This helps to ensure that the systems are not over-tuned to specific summary lengths. For each summary task (i.e., all document sets and all summary lengths), there are three distinct gold standard summaries created by different human analysts.

Pre-processing includes the same modules that are used for the relation extraction corpora (described in Chapter 3). This includes sentence boundary identification, segmentation of words (tokenisation), labelling words with part-of-speech tags,

[8] http://www-nlpir.nist.gov/projects/duc/index.html

identification of noninflected base word forms (lemmatisation) from the LT-TTT tools (Grover et al., 2000). Pre-processing also includes dependency parsing using Minipar (Lin, 1998) and automatic named entity recognition using the C&C tagger (Curran and Clark, 2003) trained on the data from the MUC-7 shared task (Chinchor, 1998). In addition to named entities, the ten most frequent nouns in each document set are marked as entities, which are marked with type label XFN. Filatova and Hatzivassiloglou (2004) introduced the tagging of frequent nouns as entities to help identify non-named referring expressions.

6.3.3 Evaluation

The evaluation uses Rouge[9] to determine which representation selects content that overlaps most with human summaries. Rouge estimates the coverage of appropriate concepts (Lin and Hovy, 2003; Lin, 2004) in a summary by comparing it to several human-created reference summaries. Rouge-1 does so by computing precision and recall based on macro-averaged unigram overlap. Rouge-SU4 does so by calculating bigram overlap where bigrams are allowed to be composed of non-contiguous words (with as many as four words intervening). Rouge-SU4 also includes unigrams to decrease the chances of zero scores where there is no skip bigram overlap.

The configuration[10] is based on comparisons between Rouge and human judgements of content coverage (Lin, 2004), which suggest that Rouge-1 and Rouge-SU4 with stemming[11] and removal of stop words are good measures for evaluating multi-document summarisation tasks, consistently achieving Pearson's correlation scores above 0.72 and as high as 0.9 for longer summaries. The results also suggest that comparison to multiple human summaries is better (especially where the number of human reference summaries is greater than or equal to three). A jackknifing procedure (i.e., $k-1$ cross-evaluation) is used here so that human gold-standard and automatic system summaries can be compared.

It has been shown in DUC 2005 and work by Murray et al. (2005b) that Rouge does not always correlate well with human evaluations. Rouge suffers from a lack of power to discriminate between systems whose performance is judged to differ by human annotators. In particular, Rouge may be misleading when comparing diverse systems

[9] Rouge stands for recall-oriented understudy for gisting evaluations. While current versions also compute precision and f-score of system summaries, the evaluation here uses recall alone. Rouge can be obtained from `http://haydn.isi.edu/ROUGE/`.

[10] i.e., `ROUGE-1.5.5.pl -n 1 -2 4 -u -m -s -r 1000 -f A -p 0.5 -t 0`

[11] Rouge stemming uses Porter's algorithm (Porter, 1980).

such as abstractive versus extractive or even when comparing two extractive systems that use different generation techniques to process the selected content. However, it is a sound measure for comparing different representations in the current evaluation, where all systems use the same sentence selection framework and the same generation techniques. Furthermore, comparison of Rouge on the data used here (DUC 2001) have shown that Pearson's correlation with human judgement is 0.73 and 0.72 for Rouge-1 and Rouge-SU4 respectively on short, 50 word summaries. On summaries of 100-400 words, both measures have higher correlations in the range of 0.83 to 0.90, generally increasing for longer summaries (Lin, 2004).

6.4 Models

In this section, the different representations compared here are described in more detail. Figure 6.5 contains an example sentence and its representation using various corresponding to the various representations of sentence content explored here.[12] These will be described in detail in the following sections.

6.4.1 Baseline *tf*idf* Representation

The baseline model represents sentences by bags-of-words using a conventional weighting scheme based on term frequency (*TF*). Following the methodology from Filatova and Hatzivassiloglou (2004), document frequencies for terms were obtained from a list of 31,928,892 terms compiled from a snapshot of 49,602,191 web pages.[13] Term weighting is calculated using *tf*idf* as:

$$w(i,j) = \sqrt{\left(1+\log\left(tf_{i,j}\right)\right) * \log\left(\frac{N}{df_i}\right)} \tag{6.2}$$

where $tf_{i,j}$ is the number of times term i occurs in sentence j and df_i is the number of documents in which term i occurs. An example sentence and its *tf*idf* representation can be seen in Figure 6.5.

[12]The sentence was selected from a document set from DUC 2001 that contains articles about Neil Bush and his role in the collapse of Silverado Savings and Loan during the U.S. Savings and Loan crisis of the 1980s and 1990s.

[13]The document frequency files were compiled from a January 2001 version of the Stanford WebBase archive (`http://dbpubs.stanford.edu:8091/~testbed/doc2/WebBase/`). The work was carried out by researchers from digital library projects at Stanford University and University of California, Berkeley. The document frequency data was downloaded 6 June 2007 from `ftp://elib.cs.berkeley.edu/outgoing/df/` which, unfortunately, is now defunct.

Example Sentence	Bush worked as an oil lease negotiator for Amoco in Denver and later started his own oil company, JNB Exploration.
*tf*idf* (*TF*)	jnb:3.55, amoco:3.13, oil:3.05, negotiator:3.04, lease:2.58, exploration:2.54, denver:2.45, bush:2.44, worked:2.28, started:2.21, later:2.13, own:1.96, company:1.94, his:1.93, as:1.56, that:1.55, an:1.54, for:1.34, in:1.33, and:1.32
event (*EV*)	<PER_bush,worked,XFN_oil>:0.00023, <PER_bush,worked,ORG_amoco>:0.00011, <PER_bush,worked,LOC_denver>:0.00011, <XFN_oil,started,ORG_jnbexploration>:0.00011, <ORG_amoco,started,ORG_jnbexploration>:0.00011, <LOC_denver,started,ORG_jnbexploration>:0.00011, <LOC_denver,started,XFN_oil>:0.00011, <ORG_amoco,started,XFN_oil>:0.00011, <PER_bush,worked,ORG_jnbexploration>:0.00003, <PER_bush,started,XFN_oil>:0.00003, <PER_bush,started,ORG_jnbexploration>:0.00003
relation (*RL*)	<ORG_amoco,rd94,LOC_denver>:0.00039, <ORG_amoco,rd505,LOC_denver>:0.00039, <XFN_oil,rd92,ORG_jnbexploration>:0.00002, <XFN_oil,rd712,ORG_jnbexploration>:0.00002, ...
ne_{event} (*EE*)	<PER_bush,XFN_oil>:0.00244, <PER_bush,LOC_denver>:0.00122, <PER_bush,ORG_jnbexploration>:0.00044, <LOC_denver,XFN_oil>:0.00033, <PER_bush,ORG_amoco>:0.00022, ...
$ne_{relation}$ (*ER*)	<LOC_denver,ORG_amoco>:0.00311, <ORG_jnbexploration,XFN_oil>:0.00155

Figure 6.5: *Example sentence and various representations of sentence content. Weights are defined in the respective model description sections.*

6.4.2 Filatova and Hatzivassiloglou Event Representation

We also compare to the atomic event (*EV*) representation from Filatova and Hatzivassiloglou (2004). As described above, this consists of $<Ent_i, Connector_j, Ent_k>$ event triples, where connectors are verbs or action nouns (i.e., nouns that are hyponyms of event or activity in WordNet) that occur in between the two entity mentions. Given a named entity recogniser and a lexical resource (WordNet), these are derived automatically from the text using a methodology that can be generalised to the framework in Figure 6.6. The algorithm processes one sentence at a time. In the first step, pairs of co-occurring entity mentions are identified. All pairs of entity mentions that occur together in a sentence are considered co-occurring at this point. Next, the algorithm characterises the entity mention pairs using event-denoting words from the intervening context, referred to as *Connectors* and discards pairs without an intervening connector word.[14]

Event triple weighting is calculated by combining entity pair and connector weights as:

$$w_{ev}(i,j,k) = w_{ne}(i,k) * w_{cn}(j,i,k) \tag{6.3}$$

where $w_{ne}(i,k)$ is the weight of the entity pair $<i,k>$ consisting of entities i and k and $w_{cn}(j,i,k)$ is the weight of connector j in the context of entity pair $<i,j>$. Filatova and Hatzivassiloglou calculate $w_{ne}(i,k)$ as the normalised entity pair count, i.e.:

$$w_{ne}(i,k) = \frac{C_{ne}(<i,k>)}{C_{ne}(<*,*>)} \tag{6.4}$$

where $C_{ne}(<i,k>)$ is the count of mentions of entity pair $<i,k>$[15] and $C_{ne}(<*,*>)$ is the total count of all mentions of entity pairs $<*,*>$. And they calculate $w_{cn}(j,i,k)$ as the normalised count of connector j in the context of the entity pair, i.e.:

$$w_{cn}(j,i,k) = \frac{C_{cn}^{<i,k>}(j)}{C_{cn}^{<i,k>}(*)} \tag{6.5}$$

where $C_{cn}^{<i,k>}(j)$ is the count of occurrences of connector j in the context of entity pair $<i,k>$ and $C_{cn}^{<i,k>}(*)$ is the total count of all connectors in the context of entity pair $<i,k>$.[16] An example sentence and its event representation can be seen in

[14] Valid connectors consist of verbs or action nouns. Action nouns are defined as nouns that are hyponyms of event or activity in WordNet. No word sense disambiguation is performed.

[15] Coreference between entity mentions is computed as based on exact string match after removing punctuation, converting to all lower case, and prefixing the entity type. For example, the entity mention string "JNB Exploration" with type ORGANISATION is normalised to `ORG_jnbexploration`.

[16] Filatova and Hatzivassiloglou do not specify whether their weighting is at the level of a single sentence or over the full document set. For the evaluation here, event weighting incorporates the two levels. This is achieved by multiplying the sentence-level weight by the document-level weight.

For each sentence:	
1	Identify event protagonists (pairs of co-occurring entity mentions)
2	Characterise event type (intervening event-denoting words)

Figure 6.6: *Filatova and Hatzivassiloglou's algorithm for atomic event extraction.*

Figure 6.5. Event triples generated for the example sentence include <PER_bush, worked, ORG_amoco> and <PER_bush, started, ORG_jnbexploration>. Some erroneous event triples are also generated, which are discussed in more detail below.

The first error has to do with the fact that entities include named entities identified in the pre-processing as well as the ten most frequent nouns in the document set. In the example sentence from Figure 6.5, non-named referring expressions include "oil lease negotiator" and "oil company" (where noun phrase heads are underlined). The most frequent nouns for this sentence's document set include 'oil' but not 'negotiator' or 'company' and the approach does not differentiate between heads and non-heads of noun phrases. Therefore, 'oil' is labelled as an entity and extracted in a number of triples such as <PER_bush, worked, XFN_oil> (as opposed to <PER_bush, worked, XFN_negotiator>).

Another problem illustrated by the example sentence has to do with the noisy nature of the surface-level approach to identifying entity pairs and connectors tends to generate many false positive events. Consider the triple <ORG_amoco, started, ORG_jnbexploration>. The sentence does not actually describe an event involving Amoco and JNB Exploration and certainly not a 'started' event between the two. Rather, it includes aspects of two events involving Neil Bush and companies that he was involved with. If the algorithm was constrained based on the underlying grammatical structure, it should be able to identity that the arguments of 'worked' are 'Bush' and 'Amoco' (i.e., <PER_bush, worked, ORG_amoco>) and that 'worked' does not describe an event involving 'Amoco' and 'JNB Exploration'.

6.4.3 GRE-based Relation Representation

The focus of the evaluation is the relation-based representation based on the GRE models developed in this thesis. Thus, relation mentions are identified using the optimised GRI approach from Chapter 4. Specifically, co-occurring entity mention pairs

are defined as all those that have either 1) no more than two intervening words in the surface order of the sentence or 2) no more than one edge intervening on the shortest path through the dependency parse. This is more strict than the Filatova and Hatzivassiloglou (2004) approach in the sense that entity mentions have to occur much closer or be connected by a single dependency relation. At the same time, it is less strict in the sense that an action- or event-denoting word is not required in the context, which reflects the focus on relations instead of events.

Relation *connectors* are derived from the output of the optimised GRC approach from Chapter 5. Specifically, the LDA model incorporating word, entity and dependency path features is used. For every relation-forming entity mention pair, this outputs a topic distribution that represents the type of relation that is described. This representation abstracts away from surface-level event descriptors used by Filatova and Hatzivassiloglou (2004) and should help to create more general models of relation types and possibly event types as well. For the purpose of comparison, relation triples are weighted in the same way as event triples using Equation 6.3 above using the same approach for entity pair weighting. The connector pair weighting is modified to use the distribution over topics given by the LDA output.[17]

An example sentence and its corresponding relation representation can be seen in Figure 6.5. Relation triples generated for the example sentence include <ORG_amoco, rd94, LOC_denver> and <ORG_amoco, rd505, LOC_denver>, where the connectors (i.e., rd94 and rd505) are identifiers that index particular topics from the LDA output. Here, rd94 and rd505 index topics that correspond to *located-in* relations so the respective triples both describe *located-in* relations between Amoco and Denver. Relation triples generated for the example sentence also include <XFN_oil, rd92, ORG_jnbexploration> and <XFN_oil, rd712, ORG_jnbexploration>. These are erroneous for the same reason as some of the event triples above, namely the fact that the shallow approach to identifying non-named referring expressions identifies 'oil' but not "oil company".

6.4.4 Entity Pair Representations

Finally, we investigate the performance of representations that do not model event or relation type information. These are identical to the EV and RL representations above,

[17] Distributions for entity mention pairs tend to have a long uniform tail and only a few topics with higher probability. In converting to a weighting scheme, topic representations here are converted to a sparse representation where all topics in the uniform tail are removed.

except they are $<Ent, Ent>$ 2-tuples instead of $<Ent, Connector, Ent>$ 3-tuples. That is, entity pairs are included here provided that they meet the GRI constraints. They are weighted using the normalised entity pair count (Equation 6.4 above). Entity pairs generated for the example sentence in Figure 6.5 include <LOC_denver,ORG_amoco> and <ORG_jnbexploration,XFN_oil>.

6.5 Experiments

6.5.1 Experiment 1: Comparing Extraction Algorithms

6.5.1.1 Method

The first experiment here compares the various greedy approaches to set cover to see which is best. Specifically, it addresses the following questions:

- *Are comparative results for extractive summarisation algorithms the same as Filatova and Hatzivassiloglou's? Are differences statistically significant?*
- *Do differences hold when using relation representation?*

This is a replication of the Filatova and Hatzivassiloglou (2004) study for the *tf*idf* representation and their *event* representation and a comparison to see if the results hold for the *relation* representation. Furthermore, this experiment seeks to establish whether any differences are statistically reliable, using paired Wilcoxon signed ranks tests across document sets.

6.5.1.2 Results

Tables 6.2 and 6.3 contain results for the *tf*idf* and *event* representations respectively. Rows in the tables represent the different extractive summarisation algorithms (described above in Section 6.3.1). The static algorithm is represented as *AL1*, the adaptive algorithm as *AL2* and the modified adaptive algorithm as *AL3*. Columns contain results for the four different lengths of summary (50, 100, 200 and 400 words). The best algorithm for each summary length is in bold. Systems that are statistically distinguishable from the best (i.e., $p \leq 0.05$) are underlined.

The results for the *tf*idf* representation demonstrate that algorithm type has no significant effect on performance, with relative ordering of the algorithms differing depending on which Rouge measure is used for all four summary lengths. The results for

Rouge-1					Rouge-SU4				
	50	100	200	400		50	100	200	400
AL1	**0.0849**	**0.1164**	0.1754	0.2438	*AL1*	**0.0221**	**0.0296**	0.0468	**0.0723**
AL2	**0.0849**	0.1102	**0.1807**	**0.2496**	*AL2*	0.0220	0.0276	**0.0476**	0.0720
AL3	0.0797	0.1113	0.1742	0.2467	*AL3*	0.0173	0.0259	0.0442	0.0693

Table 6.2: *Comparison of extractive summarisation algorithms for the* tf*idf *representation. Rows correspond to the static (*AL1*), adaptive (*AL2*) and modified adaptive (*AL3*) algorithms. The best score in each summary length column is in bold and those that are statistically distinguishable from the best are underlined.*

Rouge-1					Rouge-SU4				
	50	100	200	400		50	100	200	400
AL1	0.1300	0.1763	0.2173	0.2821	*AL1*	0.0358	**0.0506**	0.0665	0.0876
AL2	**0.1412**	**0.1778**	0.2237	0.2899	*AL2*	**0.0397**	0.0499	0.0679	0.0898
AL3	0.1360	0.1776	**0.2315**	**0.3019**	*AL3*	0.0376	0.0494	**0.0692**	**0.0950**

Table 6.3: *Comparison of extractive summarisation algorithms for the* event *representation. Rows correspond to the static (*AL1*), adaptive (*AL2*) and modified adaptive (*AL3*) algorithms. The best score in each summary length column is in bold and those that are statistically distinguishable from the best are underlined.*

Rouge-1					Rouge-SU4				
	50	100	200	400		50	100	200	400
AL1	0.1484	0.1645	0.2049	0.2718	*AL1*	0.0430	0.0464	0.0610	0.0869
AL2	**0.1525**	0.1626	0.2117	0.2664	*AL2*	**0.0459**	0.0443	0.0615	0.0810
AL3	0.1360	**0.1766**	**0.2412**	**0.3014**	*AL3*	0.0356	**0.0491**	**0.0701**	**0.0939**

Table 6.4: *Comparison of extractive summarisation algorithms for the* relation *representation proposed here. Rows correspond to the static (*AL1*), adaptive (*AL2*) and modified adaptive (*AL3*) algorithms. The best score in each summary length column is in bold and those that are statistically distinguishable from the best are underlined.*

the *event* representation suggest that algorithm type has a small effect on performance. For summaries of length 50 evaluated with Rouge-1, the adaptive algorithm (*AL2*) is significantly better than the static algorithm (*AL1*). And, for summaries of length 400 evaluated with Rouge-1, the modified adaptive algorithm (*AL3*) is significantly better than the static algorithm (*AL1*). This is a weak result but is generally in line with Filatova and Hatzivassiloglou's conclusion that an improvement may be achieved by using adaptive algorithms with the *event* representation. On the other hand, this does not support their conclusion that the modified adaptive algorithm also leads to improvements over the static algorithm when using the *tf*idf* representation.

Table 6.4 contains results for the different extractive summarisation algorithms using the *relation* representation. Here, there is a stronger effect of algorithm type on system performance. The modified adaptive algorithm (AL3) scores significantly better than the other algorithms when evaluating summaries of length 200 and 400 with Rouge-1. And, it scores significantly better than the other algorithms when evaluating summaries of length 400 with Rouge-SU4. Based on these results, the modified adaptive algorithm is used for all of the remaining experiments.

6.5.2 Experiment 2: Relation-Based Representations

6.5.2.1 Method

The second experiment looks at the potential contribution of *relation* representations for extractive summarisation. It addresses the following questions:

- *Which is the best GRC configuration for summarisation using relation representations?*
- *Can extractive summarisation be improved using representations based on generic relation extraction? How do relation representations compare to event representations?*

This seeks to determine which of the various approaches to relation characterisation is most beneficial to extractive summarisation. This also evaluates the contribution of *relation* representations with respect to conventional *tf*idf* representations and compares this to the related improvements using Filatova and Hatzivassiloglou's *event* representation. Based on results from the previous section, the modified adaptive algorithm (*AL3*) is used here.

6.5.2.2 Results

Table 6.5 contains results for various *relation* representations based on different approaches to GRC. Rows in the table represent the different feature sets used for relation characterisation. The feature set consisting of word and dependency features is represented as WD, the feature set consisting of entity and dependency features as ED, and the feature set consisting of all three feature types as WED. Free parameters are tuned individually for each feature set on the ACE data (described in Chapter 5). Columns contain results for different lengths of summary (50, 100, 200 and 400 words). As above, the best feature set for each summary length is in bold and feature sets that are statistically distinguishable from the best (i.e., $p \leq 0.05$) are underlined. The results show that the full feature set (WED) achieves the highest mean Rouge scores for all summary sizes. Furthermore, the WED feature set is always significantly better than the WD feature set though it is not always significantly better than the ED feature set. These results suggest that the entity features are important despite the fact that entities are explicitly modelled in $<Ent_i, Connector_j, Ent_k>$ triples. This confirms that the best feature set from the GRC experiments in Chapter 5 is also the best feature set for the extrinsic summarisation task.

Rouge-1					Rouge-SU4				
	50	100	200	400		50	100	200	400
WD	0.1086	0.1493	0.2068	0.2823	WD	0.0267	0.0364	0.0556	0.0844
ED	0.1192	**0.1766**	0.2171	0.2932	ED	0.0288	0.0454	0.0618	0.0908
WED	**0.1360**	**0.1766**	**0.2412**	**0.3014**	WED	**0.0356**	**0.0491**	**0.0701**	**0.0939**

Table 6.5: *Comparison of Rouge scores for different relation representations. Rows correspond to the different GRC feature combinations based on intervening words (*W*), entity words (*E*) and dependency paths (*D*). The best score in each summary length column is in bold and those that are statistically distinguishable from the best are underlined.*

Rouge-1					Rouge-SU4				
	50	100	200	400		50	100	200	400
TF	0.0797	0.1113	0.1742	0.2467	*TF*	0.0173	0.0259	0.0442	0.0693
EV	**0.1360**	**0.1776**	0.2315	**0.3019**	*EV*	**0.0376**	**0.0494**	0.0692	**0.0950**
RL	**0.1360**	0.1766	**0.2412**	0.3014	*RL*	0.0356	0.0491	**0.0701**	0.0939
HU	0.4870	0.5197	0.5655	0.6045	*HU*	0.3803	0.3990	0.4159	0.4324

Table 6.6: *Comparison of Rouge scores for the* tf*idf *(*TF*),* event *(*EV*) and* relation *(*RL*) representations with respect to the human upper bound (*HU*). The best score in each summary length column is in bold and those that are statistically distinguishable from the best are underlined.*

Table 6.6 contains results for *tf*idf* and *relation* representations. It also contains results for the human upper bound obtained by comparing the gold standard summaries with each other (described in Section 6.3.3). Here, the *tf*idf* representation is referred to as *TF*, the *event* representation as *EV*, the *relation* representation as *RL* and the human upper bound as *HU*. The results demonstrate unambiguously that the *event* and *relation* representations outperform the *tf*idf* representation, with strongly significant p-values less than 0.001 for both Rouge measures and all summary lengths. The *event* and *relation* representations are indistinguishable for both Rouge measures and all summary lengths. Finally, there is still room for dramatic improvement as both of the *event* and *relation* representations are much lower than the human upper bound.

6.5.3 Experiment 3: Contributions of GRI and GRC

6.5.3.1 Method

The third experiment isolates the effect of generic relation identification. It addresses the following question:

- *How does the entity pair representation based on the GRI models developed in Chapter 4 perform with respect to the entity pair representation based on Filatova and Hatzivassiloglou event extraction?*

- *How do the event and relation representations perform with respect to corresponding entity pair representations?*

This serves as an evaluation of the optimised generic relation identification without characterisation. This also serves as a comparison to a similar though much simpler representation that could be considered a strong baseline.

6.5.3.2 Results

Table 6.7 contains results for the entity pair representations (described in Section 6.4.4). Rows contain the different representations based on the approaches to entity pair identification for relations (*ER*) and events (*EE*) respectively. Columns contain results for different lengths of summary (50, 100, 200 and 400 words). The best representation for each summary length is in bold and representations that are statistically distinguishable from the best (i.e., $p \leq 0.05$) are underlined. In contrast to the results for normal *tf*idf*, *relation* and *event* representations which use the modified adaptive algorithm, all results for entity pair representations use the adaptive algorithm.[18] The results suggest that the entity pair model that has been optimised on relation extraction data outperforms the entity pair model based on Filatova and Hatzivassiloglou's event extraction algorithm, at least for medium sized summaries of 100 and 200 words where *ER* is significantly better than *EE* for both Rouge measures.

The scores for the entity pair representations reported in Table 6.7 are statistically indistinguishable from those for the corresponding *relation* and *event* representations in Tables 6.3 and 6.4 above. This appears to be a mixed result for both the relation

[18]For *ER*, the adaptive algorithm is significantly better than the static and modified adaptive algorithms ($p \leq 0.01$) for both Rouge measures and all summary lengths. Rouge-SU4 results are not significant but have the same trend.

Rouge-1					Rouge-SU4				
	50	100	200	400		50	100	200	400
ER	**0.1497**	**0.1929**	**0.2527**	**0.3123**	*ER*	**0.0419**	**0.0537**	**0.0786**	**0.1008**
EE	0.1442	0.1705	0.2288	0.3061	*EE*	0.0364	0.0447	0.0643	0.0963

Table 6.7: *Comparison of Rouge scores for entity pair representations based on relations (*ER*) and events (*EE*). The best score in each summary length column is in bold and those that are statistically distinguishable from the best are underlined.*

representation introduced here and the Filatova and Hatzivassiloglou event representation. While optimised relation identification is shown to have a positive effect on Rouge scores when compared to relation identification based on Filatova and Hatzivassiloglou's approach to atomic event extraction, the same cannot be said of approaches to characterising relation and event types. However, as the correlation analysis (Section 6.6.1 below) demonstrates, RL and ER do not necessarily perform well on the same document sets. This suggests that they are actually complementary to some degree meaning that a combined systems based on both representations would outperform RL and ER on their own.

6.6 Analysis

6.6.1 Complementarity

Figure 6.7 contains results for a correlation analysis comparing the various representations: *tf*idf* (*TF*), *event* (*EV*), *relation* (*RL*), entity pair based on *event* extraction (*EE*) and entity pair based on *relation* extraction (*ER*). This also includes a comparison to the human upper bound (*HU*). Values are arranged in a matrix where cells contain the association values measured across document set Rouge-SU4 scores[19] using Spearman's ρ rank correlation coefficient (r_S). Here, high values mean that two representations tend to perform well on the same document sets such that an ordering of document sets by Rouge scores is similar for the representations being compared. In the figure, association strength is represented by shading where light-toned squares indicate strong correlation (and the darkest squares indicate weak negative correlation). For example,

[19] Association across document set Rouge-1 scores shows similar trends.

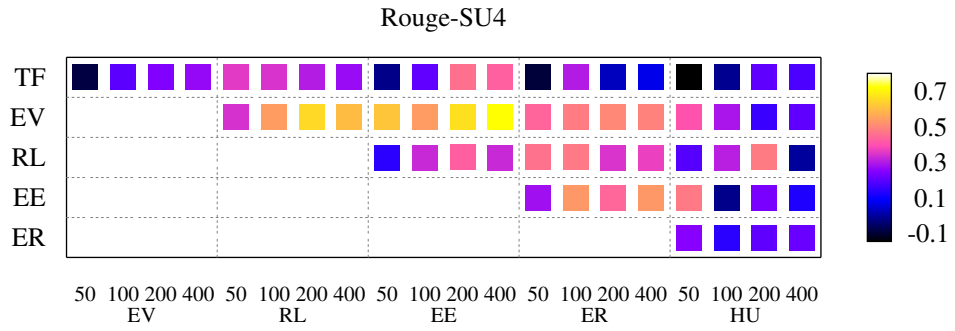


Figure 6.7: *Comparison of representations using Spearman's r_s. Row and column labels correspond to* tf*idf *(*TF*),* event *(*EV*),* relation *(*RL*), event entity pair (*EE*), relation entity pair (*ER*) and human (*HU*) representations. Lighter toned squares indicate stronger correlation.*

the upper left cell contains Spearman's r_S between the TF and EV representations. The four squares correspond to r_S values of -0.085, 0.199, 0.245 and 0.267 respectively for summaries of 50, 100, 200 and 400 words.

The analysis illustrates a number of interesting points. First, it demonstrates that none of the representations correlate highly with the human upper bound. This means that the automatic systems do not necessarily do well on the document sets that may be considered easier as measured by human agreement using Rouge. Thus, it suggests that task difficulty (as measured by human agreement) does not need to be considered as a possible underlying cause of correlation between the automatic systems. The analysis also illustrates that there is no clear and consistent relationship between summary length and correlation values. Some cells suggest that correlation may have a monotonic linear relationship increasing with length (e.g., TF*EV) while others seem to suggest inverse linear (e.g., TF*RL), quadratic (e.g., EV*HU) and invariant (e.g., EV*EE) relationships with length.

Looking at correlation between automatic systems, correlation values closer to zero suggest that the systems do well on different document sets and that a combined system might therefore be better. By this reasoning, the largest gains would come from combining TF with any other representation. Among the other automatic systems, the *relation* representation (RL) shows moderately strong potential for combination with its corresponding entity pair representation (ER) with Spearman's r_S values in the range from 0.348 to 0.476. This suggests that ER should not necessarily be considered

a simpler representation of the same information captured by RL when comparing results. The *event* representation (EV), by contrast, shows the weakest correlation of any comparison with its corresponding entity pair representation (EE) with r_S values in the range from 0.541 to 0.725.

6.6.2 Error Analysis

This section addresses the types of errors that characterise the automatic summarisation systems. Figures 6.8 and 6.9 contain summaries from document sets where the *relation* (*RL*) and *event* (*EV*) representations perform poorly with respect to the *tf*idf* (*TF*) representation. Figures 6.10 and 6.11 contain summaries from document sets where *RL* performs well with respect to *EV* and/or *TF*. The figures also contain human gold standard summaries (*HU*).[20] The numbers in parentheses below each summary representation label indicate the Rouge-SU4 score and the rank of document set rank for the representation according to the Rouge-SU4 scores. Automatic system summaries should be viewed as a list of extracted sentences due to the fact that the evaluation here is exclusively concerned with content and no attempt is made to address coherence. The bold numbers in square brackets before each sentence indicates its rank in terms of the order in which it was extracted. If part of the final sentence was shortened for evaluation so as not to exceed the word limit, then this is indicated by the word "END" in square brackets.

The document set for the summaries in Figure 6.8 illustrate a situation where the *relation* and *event* representations performed worse than *tf*idf*. The gold standard summary describes a disease outbreak event, addressing the cause, affected regions and susceptible population as regards the increase in cases of tuberculosis in the 1980s and 1990s. The particularly poor performance of the *event* representation seems to be due to the fact that the sentences selected contain a very large number of false positive events. The third sentence, the most extreme example, contains 176 events. These include the true positive atomic events like <ORG_centers, recorded, XFN_cases>, but it also includes many false positive atomic events such as <LOC_wyoming, recorded, XFN_cases>, <XFN_tb, is, PER_leebreichman> and <PER_leebreichman, division, LOC_dentistry>. The false positive atomic events are due primarily to the event identification algorithm being noisy on sentences with many entity mentions and many valid connectors. The sentence under considera-

[20] There are three human summaries for each document set. The summary presented here is the one created by the same individual who chose the topic and collected the documents.

TF (0.046) (3/29)	**[S1]** The CDC committee also recommends new tests at least once a year, rapid chest X-rays for TB-infected people showing symptoms, and isolation – off the prison property, if necessary – for those with suspected or confirmed symptomatic TB cases. **[S2]** At the present time, there are no laws in Latvia that would require infectious cases of tuberculosis to undergo mandatory treatment. **[S3]** YOU'RE INTERESTEDEmployees who think that poor air quality at their workplace is contributing to health problems can call Cal-OSHA, a state agency that monitors worker safety and health. **[S4]** Moreover, unlike AIDS, TB is a highly contagious **[END]** disease that can be spread by airborne particles coughed up by a person with untreated, clinically active pulmonary TB.
EV (0.023) (23/29)	**[S1]** Those at high risk for contracting TB are people whose capacity for resisting infection is weakened, either through diseases such as HIV infection, by drug or alcohol abuse, serious illness such as cancer, or by poor nutrition. **[S2]** Stead's study, published in the New England Journal of Medicine, was based largely on a review of 25,398 elderly people who were free of TB infection when they were admitted to Arkansas nursing homes. **[S3]** In Wyoming, the Centers for Disease Control recorded no new cases of TB in 1989, demonstrating that it is a condition that can be controlled and cured **[END]**, according to Dr. Lee B. Reichman, director of the pulmonary division of the University of Medicine and Dentistry in New Jersey.
RL (0.035) (17/29)	**[S1]** Twenty-three percent of those with HIV also carried TB bacteria as did 20 percent of those who were free of the AIDS virus. **[S2]** AIDS tests should be offered to all inmates with known TB infections, the CDC report said. **[S3]** "In some large correctional systems, the incidence of TB has increased dramatically," the CDC said, noting that in New York state there were 106 TB cases per 100,000 inmates in 1986 – seven times more than the average of 15 cases reported in 1976-78. **[S4]** In homes where the initial source of the disease was white, 17 percent of **[END]** blacks and 12 percent of whites caught the infection.
HU (0.354) (26/29)	The occurrences of tuberculosis increased in the 1980s after a three decade decline. By 1990 it was the world's deadliest infectious disease, killing three million annually. The tuberculosis was fueled by AIDS patients who were vulnerable when their lowered immune system allowed the latent bacteria to develop into active tuberculosis. They then transmitted it to others. Tuberculosis ran rampant in sub-Saharan Africa, and increased in Latin America and Southeast Asia. In the United States the highest rates of infection were in the Northeast. Prisoners are highly susceptible to the disease. Airtight buildings with bad ventilation spawns tuberculosis. **[END]**

Figure 6.8: *Example system and* human *(*HU*) summaries where the* relation *(*RL*) and* event *(*EV*) representations perform poorly with respect to the* tf*idf *(*TF*) representation: Tuberculosis Document Set (d15).*

TF (0.033) (7/29)	**[S1]** Seven thousand British and 4,000 French workers, on eight- to 12-hour shifts, have completed more than half of the job. **[S2]** But the men staying here – most of them migrant laborers from Ireland and northern England – are rumored to earn as much as $1,750 a week. **[S3]** The French favourite is Le Touquet, still an elegant place for Parisians to spend le weekend. **[S4]** Critics say the tangle of commuter lines in south-east England, so obsolete that trains can be delayed by a sudden fall of autumn leaves, will delay tunnel traffic. **[S5]** Mr Jo Libeer, managing director of the Courtrai **[END]** chamber of commerce, is equally optimistic about the likely impact on the area of the tunnel.
EV (0.034) (16/29)	**[S1]** Eurotunnel will run shuttle trains once every three minutes at peak times between terminals near Folkestone and Calais, and British Rail and the Frenh state railroad will operate trains from London and Paris. **[S2]** Giant boring machines are digging three tunnels toward each other from Folkestone, England and Calais, France, with the first underground meeting expected in November in the service tunnel between the rail tunnels. **[S3]** President Francois Mitterrand and Mrs. Thatcher are expected to meet each other in the tunnel Jan. **[S4]** Once the tunnel is open, said Parry, industry will be attracted to the area and people will move in. **[END]**
RL (0.014) (28/29)	**[S1]** When Kelly sees Range Rover and Jaguar drivers collecting their cases of wine in Hesdin, she would like them to drop into her office 100 metres away and choose a house as well. **[S2]** Many observers believe that, by doing nothing to improve the nation's creaking rail system, Britain will lose out on the full benefits of the Chunnel. **[S3]** Critics say this will waste time and Britain should follow the continental practice of handling such matters on the train during the journey. **[S4]** Eurotunnel will run shuttle trains once every three minutes at peak times between terminals near Folkestone and Calais **[END]**, and British Rail and the Frenh state railroad will operate trains from London and Paris.
HU (0.392) (16/29)	British and French workers were expected to complete the 31-mile "chunnel" between England and France by May 1994. It would cut the London-Paris journey from six hours to three and reinforce trade. Initially, the English showed that they did not want a fixed link, fearing rabies, rats and terrorists coming through. In rural SE England, they were reluctant to have high-speed trains screaming through. The French viewed the chunnel positively, expecting it to revitalize its depressed northern regions where the tunnel surfaces and as buyers and entrepreneurs set up bases on the Continent, and holidaymakers and freight **[END]** traffic heads for the tunnel.

Figure 6.9: *Example system and* human *(*HU*) summaries where the* relation *(*RL*) and* event *(*EV*) representations perform poorly with respect to the* tf*idf *(*TF*) representation: Channel Tunnel Document Set (d39).*

tion has nine entity mentions (LOC_wyoming, ORG_centers, ORG_diseasecontrol, XFN_cases, XFN_tb, PER_leebreichman, ORG_universityofmedicine, LOC_dentistry, LOC_newjersey) and nine connectors (recorded, demonstrating, is, be, condition, controlled, cured, according, division). Some of the false positive errors in this sentence can also be attributed to bad named entity recognition. For example, "Centers for Disease Control" is incorrectly recognised as two separate entity mentions (i.e., "Centers" and "Disease Control", normalised respectively to ORG_centers and ORG_diseasecontrol). More generally, the *event* and *relation* representations for this evaluation cannot capture date, time or numeric event or relation information as these entity types have not been included. Neither do they capture event or relation information involving non-MUC entity types such as disease names (e.g., "tuberculosis"). Finally, the *event* and *relation* representations do not capture descriptive or analytic information such as the details about susceptible populations.

The document set for the summaries in Figure 6.9 illustrates a second situation where the *relation* and *event* representations performed worse than *tf*idf*. However, here it is the *relation* representation that does particularly poorly. The gold standard summary describes a construction event, addressing the building of the Channel Tunnel with a focus on differing attitudes towards the project in Britain and France. The poor performance of the *relation* representation seems to be due in part to the fact that it gives undue importance to irrelevant and noisy relationships, e.g. between PER_kelly and ORG_rangerover and between ORG_rangerover and ORG_jaguar in the first sentence. More generally, the relatively poor performance of the *relation* and *event* representations is due to the fact that they do not capture sentiment, which is the main focus of the summary.

The document set for the summaries in Figure 6.10 illustrates a situation where the *relation* and *event* representations perform well with respect to *tf*idf*. The gold standard summary describes a beating event, addressing the basic facts of the Rodney King beating by Los Angeles police as well as the political aftermath which consists primarily of a related investigation event and a summary of related police brutality events. The difference in performance seems to be due to the fact that relations and events are central to all aspects of this summary and the *relation* and *event* representations clearly do better than *tf*idf* at capturing this information. This summary also illustrates an unintended side-effect of the *relation* representation where the generic relation identification algorithm finds relations between components of lexical compounds or multi-word phrases. The representation for the third sentence in the *RL*

summary, for example, includes a relation between ORG_police and XFN_chief in addition to true positive relations e.g. between ORG_police and PER_darylgates and false positive relations e.g. between PER_tombradley and ORG_police.

The document set for the summaries in Figure 6.11 illustrate a second situation where the *relation* representation performs well. However, here it is the *event* representation that performs poorly by comparison. The gold standard summary is a biographical sketch of a political group, addressing relations like the leaders, location, constituency and rivals of the the Peru's Shining Path Maoist group as well as events like founding and leadership succession. The fact that the *relation* representation outperforms the *event* representation here seems to be due to the fact that relations are more important than events in this document set. The representation for the first sentence of the *EV* summary, for example, has 63 events consisting of various combinations of the sentence's entities (ORG_bup, ORG_centralcommittee, PER_feliciano, ORG_newpower, ORG_peoplesliberationarmy, and ORG_blackgroup) and connectors (presided, orders, movement, order, party, based, break). The representation for the first sentence of the *RL* summary, by contrast, contains three relations between ORG_maoistshiningpath and XFN_guerrillas, between LOC_peru and LOC_huallagarivervalley and between XFN_guerrillas and LOC_peru.

This error analysis suggests that the different approaches here are appropriate for different types of summary task. The *relation* and *event* representations perform poorly on summarisation tasks that are oriented towards sentiment, description or analysis. However, they do well on document sets that are oriented towards relation and event information typical to information extraction tasks. This supports the notion from the previous section that the different representations evaluated here are complementary.

6.7 Summary and Future Work

This chapter presented results of an extrinsic evaluation of generic relation extraction, demonstrating that it is an effective representation for sentence extraction for multi-document summarisation. The *relation* representation was compared to a non-trivial *tf*idf* baseline and found to perform significantly better for a range of summary lengths. Related representations based on events and entity pairs exhibited statistically indistinguishable performance. A correlation analysis suggested that different representations are complementary due to the fact that they perform well on different document sets. Error analysis supported this conclusion, suggesting that the *relation* and

TF (0.016) (20/29)	**[S1]** Mr. Williams likened the report to the Knapp Commission, a 1970s blue-ribbon study that exposed widespread corruption in the New York Police Department and led to significant improvements there. **[S2]** "There's no doubt in our mind that the only reason they stopped Joe Morgan was because he is black and he was the first black who happened to come by," said William Barnes, one of the attorneys representing the former ballplayer. **[S3]** Joseph McNamara, retired chief of San Jose's department and now a fellow at Stanford University's Hoover Institution, said he has been getting calls all summer from **[END]** cities around the country about racism and brutality in their departments.
EV (0.060) (9/29)	**[S1]** A high-ranking commission appointed after the beating, under the chairmanship of Mr Warren Christopher, a lawyer and former deputy secretary of state, concluded that the Los Angeles police department got results, in terms of arrests, but had developed a 'siege mentality that alienates the officer from the community'. **[S2]** The images of Los Angeles police swinging nightsticks at King as he lay on the ground, played repeatedly on national news programs, were burned into the national conscience and led to widespread calls for investigation of police brutality. **[S3]** Besides recommending that Mr Gates should go, the Christopher commission urged a policy **[END]** of community policing with more foot patrols, as well as measures to discipline racist police officers and to improve the investigation of complaints about police brutality.
RL (0.094) (3/29)	**[S1]** Mr. Gates opposed the Police Corps because its members would not be professionals. **[S2]** Shortly after Rodney King's beating, a news program on ABC illustrating police brutality showed a still photo of police using a martial-arts weapon against a person being arrested, but there was no mention that the episode involved Operation Rescue. **[S3]** The report was issued yesterday by a commission appointed by Mayor Tom Bradley and Police Chief Daryl Gates in the wake of the videotaped beating March 3 of a black motorist, Rodney King, by Los Angeles police. **[S4]** Investigations have been launched by the FBI, the Los **[END]** Angeles County district attorney's office and the Long Beach Police Department.
HU (0.400) (15/29)	The most important of the many cases of police brutality reported in southern California 1989-1992, was the beating of Rodney King by four Los Angeles officers on March 3, 1991. An investigating commission outlined steps for improvement of the police department and called for the resignation of Chief Gates. Gates did not resign until the following year after the acquittal of the four officers caused massive rioting. Other cases of police brutality arose in Minneapolis, Chicago and Kansas City. Operation Rescue claimed that its non-violent anti-abortion demonstrators were seriously injured by excessive police tactics in more than **[END]** 50 cities.

Figure 6.10: *Example system and* human *(*HU*) summaries where the* relation *(*RL*) and* event *(*EV*) representations perform well with respect to the* tf*idf *(*TF*) representation: Police Brutality Document Set (d06).*

TF (0.035) (6/29)	**[S1]** (Note 3) For example, in his presentation of the video recording of Guzman's first call for peace talks, Fujimori claimed that the Shining Path "political" leadership "has tacitly admitted that the Peruvian state has totally recovered the initiative in confronting the Shining Path" (Lima Radio and TV, 4 October 1993). **[S2]** During the 1992 partisan meeting, it was stated that during an adverse situation the party should draft "a new plan, taking into account the experience of the past years, establish new axes, sub-axes, guidelines, and lines of action with a nationwide criteria (. . .), seek new ways to develop and **[END]** set up strategic military plans, and establish, for example, those objectives and carry them out on an established date."
EV (0.020) (25/29)	**[S1]** The Central Committee Plenum that "Feliciano" presided over in order to break away from the Black Group orders the "unleashing of a massive reassertion movement based on the BUP throughout the party, the People's Liberation Army, and among the masses of the New Power." **[S2]** The second point of the document of the Central Committee meeting presided over by "Feliciano" highlights the agreements of the Working Meeting of the Shining Path leadership held in August 1993, almost one year after Guzman was arrested, and during which the implementation of the agreements of the Third Plenum held in March 1992 was **[END]** discussed.
RL (0.078) (5/29)	**[S1]** The Maoist Shining Path guerrillas who dominate Peru's Upper Huallaga River Valley have brought their own law and order to a cocaine-corrupted, violence-ridden region. **[S2]** – Also in response to the second letter, Caretas on 14 October claimed that Fujimori was "using" Guzman "in the campaign" to win the 31 October referendum, noting that "with Guzman's letters" calling for peace talks, Fujimori's "promise" to wipe out the Shining Path by 1995 "may gain credibility." **[S3]** Police captured the top military leader of the Shining Path, a Maoist rebel group whose eight-year guerrilla war has taken more than **[END]** 10,000 lives in Peru, officials said Monday.
HU (0.363) (23/29)	The Shining Path is a Peruvian Maoist group founded in 1970 by Abimael Guzman. This group sought to overthrow Peru's elected government and install a peasant and worker state. For 10 years is worked among the Indians in the Andes, then began guerrilla operations eventually moving into the jungles and cities. Until Fujimori was elected president in 1990, Peruvian presidents had not been successful in handling the Shining Path, but, by 1994, many leaders of the group, including Guzman, were in jail. A faction of free guerrillas also had formed and was beginning to assert power repudiating Guzman and **[END]** his negotiations with the government.

Figure 6.11: *Example system and* human *(*HU*) summaries where the* relation *(*RL*) and* tf*idf *(*TF*) representations perform well with respect to* event *(*EV*) representation: Shining Path Document Set (d53).*

event representations perform poorly on summarisation tasks that are oriented towards e.g. sentiment, description or analysis while they perform well on tasks that focus on fact-oriented information typical to information extraction tasks.

Given the complementarity of representations suggested by the analysis, better performance might be obtained by combining representations. One simple approach would be to convert extraction scores from the various representations to ranks, which could be simply combined by taking the mean. There may be better combination metrics though. A user evaluation of the summary output could be used to further examine the hypothesis that different representations are preferable for different types of summaries and provide more detailed criteria for combination. The summarisation approach could also be extended to incorporate query- or topic-relevance into the extraction scores and be incorporated into a question answering system to address information requests that require information about relations between entities.

Chapter 7

Conclusion

Relation extraction is a highly promising technology for converting unstructured text data into a format that can be more effectively used for querying and automated reasoning. However, adapting conventional systems to new domains, tasks or languages requires significant effort from annotators and developers. This thesis addresses the adaptation problem by applying generic techniques that achieve comparable accuracy when transferred, without modification of model parameters, across domains and tasks. This chapter contains a summary discussion of thesis outcomes and directions for future work.

7.1 Primary Outcomes

While relation extraction promises to be a powerful technology for extracting structured information from unstructured text data, conventional approaches incur development costs that are often prohibitively expensive. In the case of rule engineering, writing extraction rules requires extensive effort from a language engineer (expert at least in rule engineering and ideally also in the target domain). In the case of supervised learning, annotation of training data and tuning features/model parameters require extensive effort from at least one annotator (expert in the target domain) and from a language engineer (expert in natural language processing). This has motivated work on partially supervised approaches for bootstrapping. However, at the least, these require seed data meaning that the relation type schema of a new application must be anticipated. This motivates the exploratory approach developed here based on generic techniques that do not require any annotation or parameter tuning when moving to new domains. This is referred to as generic relation extraction (GRE).

This thesis contributes a unified approach to GRE that synthesises the previously

disparate literatures on relation mining and relation discovery. Primary outcomes include new state-of-the-art approaches to relation identification and characterisation that incorporate governor-dependency information. Dimensionality reduction is also introduced as a step in building similarity models for the relation characterisation clustering task. Use of the ACE 2005 data, which contains markup from two annotators and an adjudicated version, allows comparison to a human upper bound for the first time. In addition, the use of gold standard relation annotation allows detailed analysis of GRE performance. This thesis also applies news-optimised models directly to a relation extraction task in the biomedical domain, demonstrating for the first time that an approach to GRE achieves comparable performance when transferred across domains. Finally, this thesis demonstrates that the relative performance of GRE models is consistent across tasks and that the GRE-based representation leads to significant improvements in sentence extraction for automatic summarisation when compared to a non-trivial baseline from the literature.

Experiments on the generic relation identification (GRI) task compared several window-based models for establishing entity mention pair co-occurrence. Combined windows based on intervening word tokens and syntactic governor-dependency paths were preferred due to significantly higher recall, which is prioritised due to the exploratory nature of the GRE task and due to the fact that applications of GRE (including the summarisation task addressed here) generally incorporate a mechanism for ranking identified relations. A correlation analysis supported this prioritisation, suggesting that ranking metrics can be used as a weak indicator of true/false relation mention status. A detailed analysis found that as much as 81% of false positive errors in the news test data (54% in biomedical data), while not marked in the gold standard, are actually implicit in the context of the sentence or document. Many of these errors would not actually be detrimental to applications of GRE. Finally, correlation analysis identified several possible indicator features that may be used as noisy filters for false positive GRI errors. These include the presence of a verb or nominalisation in the intervening window context. They also include the presence of a conjunction or disjunction.

Experiments on the generic relation characterisation (GRC) task compared similarity models for clustering entity mention pairs by relation type. Novel feature sets based on information from governor-dependency paths were shown to lead to improvements, as was the introduction of dimensionality reduction. Comparison of dimensionality reduction techniques showed that a model using latent Dirichlet allocation (LDA) – a probabilistic generative approach – successfully incorporates a larger and more in-

terdependent feature set than an unreduced model and a linear algebraic model based on singular value decomposition (SVD). The LDA-reduced system does particularly well in terms of recall. This is attributed to the LDA hyperparameters, which control the impact of sparsity. Analysis suggests that false positive errors tend to coincide with relation mentions that require inference or use figurative language. In addition, a number of errors were deemed to constitute subtle differences between gold standard relations types that are not essential to applications of GRE. False negative errors tend to coincide with relation mentions that have bad dependency paths, require inference, use figurative language or contain transcription errors. While SVD may be preferred in other application scenarios, LDA is preferred here due to accuracy and interpretability.

Finally, experiments on extractive multi-document summarisation explore GRE relation output as a novel knowledge stream for interpretation and representation of source documents. This serves as an extrinsic evaluation of end-to-end GRE based on the models developed in this thesis, demonstrating a significant improvement over a non-trivial *tf*idf* baseline. This also shows that the approaches to GRI and GRC developed here are capable of generalising across tasks. Detailed analysis suggested that the different representations compared are complementary. Specifically, representations based on relations and events tend to perform poorly on tasks that are oriented towards e.g. sentiment, description or analysis while they perform well on tasks that focus on factual information. This complementarity of representations suggests that better performance might be obtained by combining representations.

Taken together, the experimental chapters of this thesis show that GRE can be improved using dependency parsing and dimensionality reduction. In addition, comparison of dimensionality reduction techniques suggests that latent Dirichlet allocation is preferable for GRE; it performs as well as or better than SVD and offers superior interpretability. Furthermore, the experimental chapters validate the claim of modification-free adaptation for the first time with respect to both domain and task. Models developed on news data are shown to have comparable results when applied directly to biomedical data and relative performance of GRE models is shown to be the same when applied to extractive summarisation.

7.2 Secondary Outcomes

In addition to the primary experimental and modelling outcomes above, this thesis also contributes to the formalisation of the GRE task. First, the GRE task is presented in a

way that unifies the previously disjoint but closely related literatures on relation mining and relation discovery. Second, evaluation data is derived from standard and publicly available materials (i.e., the ACE 2004, ACE 2005 and BioInfer data sets), making it possible to replicate experiments. A three-stage process (re-factoring, pre-processing, re-annotation) was described for adapting these corpora to the GRE task. And, a standard XML document type for marking relation extraction data as token offsets was proposed. Finally, a detailed framework was presented for development and evaluation testing with respect to the gold standard relation extraction data. This provided for a rigorous experimental design with held-out evaluation data sets in multiple domains and the use of paired Wilcoxon signed ranks tests to quantify significant differences across entity pair sub-domains.

7.3 Future Work

This thesis only scratches the surface of research on and applications of the GRE task. Some future work was mentioned in the experimental chapters with respect to further exploration of models. This included indicator features for filtering GRI false positive errors such as verbal connector words or conjunctions on the intervening dependency path. It also included the exploration of relation extraction pattern models from Greenwood and Stevenson (2007) as a possible means of improving accuracy of GRI based on topic-focused document sets returned from IR queries. With respect to GRC, there are many options for extending models due to the flexible nature of probabilistic topic modelling. Some possibilities include: 1) non-parametric models where the number of topics is sampled (e.g., Blei et al., 2004; Teh et al., 2004), 2) variations of the model topology that model topics with respect to entity mention pairs, akin to author-topic and author-recipient-topic models (e.g., Rosen-Zvi et al., 2004; McCallum et al., 2004), and 3) approaches that integrate coreference into the GRE task. Coreference information could be integrated by augmenting the feature space based on the output of a preceding coreference module, akin to related work in summarisation (e.g., Steinberger et al., 2005). Coreference information could also be integrated using joint or iterative models that use coreference information to inform relation extraction models and vice versa akin to related work in named entity recognition (e.g., Wellner et al., 2004). With respect to the latter approach, one could use distributional information over relation types and related entities to contribute to similarity models for entity mention coreference.

Another direction for future research lies in the combination of generic approaches and semi-supervised bootstrapping to create a completely bottom-up approach. This would use the GRE models developed in this thesis to initialise the semi-supervised bootstrapping approaches. Active learning could be initialised this way by identifying pairs of clustering instances that lie at decision boundaries (e.g., maximally dissimilar pairs within a cluster or maximally similar pairs across clusters) to be presented to human annotators. Annotators could choose to introduce pairwise constraints requiring that these be in the same cluster or not in subsequent partitions. Pairwise constraints can be incorporated using semi-supervised clustering approaches based on learnt similarity measures (e.g., Klein et al., 2002; Xing et al., 2003; Bilenko et al., 2004) or by assigning arbitrarily large or small similarity values to pairs annotated respectively as having or not having the same type (e.g., Blum and Chawla, 2001). Other bootstrapping approaches could also be initialised this way, such as the iterative approaches that are seeded by example entity pairs of a specific type of relation (e.g., Agichtein and Gravano, 2000; Tomita et al., 2006). Instead of manually seeding, these approaches could be initialised by choosing instances that are representative of a certain cluster (e.g., instances that are close to the cluster centroid) and using them to induce new extraction rules and subsequently identify more example entity pairs.

In addition, various extrinsic tasks have been mentioned in this thesis to motivate the utility of the GRE task. The multi-document summarisation task demonstrated the extrinsic utility of the GRE models developed here. However, the experimental results and analysis suggested that better performance could be obtained by combining representations. One simple approach would be to convert extraction scores from the various representations to ranks, which could be simply combined by taking the mean. There may be better combination metrics though. A user evaluation of the summary output could be used to further examine the hypothesis that different representations are preferable for different types of summaries and provide more detailed criteria for combination. The summarisation approach could also be extended to incorporate query- or topic-relevance into the extraction scores and be incorporated into a question answering system to address information requests that require information about relations between entities.

The GRE-based approach to generating entity sketches (Chapter 3) could also be further explored. This could be incorporated into abstractive summarisation as a method for creating knowledge bases for natural language generation (e.g., White et al., 2001). It could also be used to identify relation triple factoids for question

answering systems. Based on manual annotation of part of the Encarta query log, Agichtein et al. (2005) show that a small number of relation types address the majority of questions that can be answered by relation factoids and thus advocate query log analysis to identify relationships most relevant to user needs. GRE could be used here as an interactive tool to help identify common relation types in query logs. Sekine (2006) suggest another approach that uses GRE as a post-processing tool to identify prevalent relations among the documents returned for a specific IR query. Yet another application of GRE-based entity sketches is the automatic generation of hyperlinks between documents. Links could be generated from a target document to other documents that describe similar relations or to other documents that describe entities that are related to (e.g., within n degrees of) entities in the target document. This could be evaluated using an information foraging task and collecting click-through data for hyperlinks and measuring the amount of time taken to perform the task.

Blue-sky directions include the automatic generation of graphic representations of entity networks that include relation types. This could follow techniques developed for the presentation of manually generated, narrative-focused networks (Lombardi et al., 2003) and incorporate time-lines to specify temporal duration of relations. Furthermore, these graphics could be hyperlinked to lead to documents describing the relations or entities. Potential applications include investigative journalism based on news archives and tax fraud detection based on filings and news archives. Related text-based automation has already been demonstrated in bioinformatics (Smalheiser and Swanson, 1998) where conventional supervised information extraction is used as a noisy approach to generating hypotheses from text data that can later be tested in wet lab experiments.

Appendix A

Document Management

A.1 An XML Document Type for RE Data

Figure A.1 contains the RE XML document type definition developed for this thesis. This includes a top-level `rexml` element containing one or more document (`doc`) elements, which are made up of a `text` element and a `markup` element. The `text` element contains the tokenised document text, marked up with in-line paragraph (`p`), sentence (`s`) and word token (`w`) information. Basic linguistic information for word tokens is encoded as attributes on `w` elements, including parts-of-speech (`p`) and lemmas (`l`). The `markup` element contains stand-off entity (`nes`) and relation (`rels`) annotation. Individual entity instances (`ne`) are specified with reference to the identifiers of the word tokens that start and end the entity text span (attributes `fr` and `to`). And individual relation instances (`rel`) are specified with reference to the entities participating in the relation (attributes `e1` and `e2`).

Figure A.2 contains an example document with the basic RE XML markup. This is drawn from the BioInfer data (see Chapter 3) and contains markup for the following sentence:

> "Beta-catenin is also found in these structures."

The markup in the figure specifies two entities (`ne`). The first `ne` element (with `id` "e75") contains the markup for the SUBSTANCE entity "Beta-catenin", which starts (`fr`) and ends (`to`) with the word token (`w`) with `id` "w211". The markup in the figure also specifies a relation (with `id` "r32") of type CAUSAL between entity "e75" ("Beta-catenin") and entity "e77" ("structures").

```
<!ELEMENT rexml (doc+)>              <!-- Rexml: Contains Doc(s)      -->
<!ELEMENT doc (text,markup)>         <!-- Doc: Contains Text, Markup -->

<!ELEMENT text (p)+>                 <!-- Text: Contains paragraphs  -->
<!ELEMENT p (s|w)+>                  <!-- P(aragraph): Contains Ss   -->
<!ELEMENT s (w+)>                    <!-- S(entence): Contains Words -->
<!ELEMENT w (#PCDATA)>               <!-- W(ord): Contains Word Text -->

<!ELEMENT markup (nes,rels)>         <!-- Markup: Contains NEs, Rels -->
<!ELEMENT nes (ne*)>                 <!-- Nes: Contains NE Mentions  -->
<!ELEMENT ne (textspan*)>            <!-- Ne: Contains NE Textspan   -->
<!ELEMENT textspan (#PCDATA)>        <!-- Textspan: Contains NE Text -->
<!ELEMENT rels (rel*)>               <!-- Rels: Contains Rel Ment'ns -->

<!ATTLIST doc id CDATA #IMPLIED>          <!-- Document ID          -->
<!ATTLIST s id CDATA #REQUIRED>           <!-- Sentence ID          -->

<!ATTLIST w id CDATA #REQUIRED>           <!-- Token ID             -->
<!ATTLIST w p CDATA #REQUIRED>            <!-- Token part-of-speech -->
<!ATTLIST w l CDATA #REQUIRED>            <!-- Token lemma          -->

<!ATTLIST ne id CDATA #REQUIRED>          <!-- NE Mention ID        -->
<!ATTLIST ne fr CDATA #REQUIRED>          <!-- NE Start Token ID    -->
<!ATTLIST ne to CDATA #REQUIRED>          <!-- NE End Token ID      -->
<!ATTLIST ne t CDATA #REQUIRED>           <!-- NE End Token ID      -->
<!ATTLIST ne st CDATA #IMPLIED>           <!-- NE Sub Type          -->

<!ATTLIST rel e1 CDATA #REQUIRED>         <!-- Rel NE 1 ID          -->
<!ATTLIST rel e2 CDATA #REQUIRED>         <!-- Rel NE 2 ID          -->
<!ATTLIST rel t CDATA #REQUIRED>          <!-- Rel Type             -->
<!ATTLIST rel st CDATA #IMPLIED>          <!-- Rel Sub Type         -->
```

Figure A.1: *Basic Document Type Definition for RE XML.*

```
<rexml>
...
<doc id='15'>
 <text>
  <p>
   <s id='s11'>
    <w id='w211' p='NN' l='beta-catenin'>Beta-catenin</w>
    <w id='w212' p='VBZ' l='be'>is</w>
    <w id='w213' p='RB'>also</w>
    <w id='w214' p='VBN' l='find'>found</w>
    <w id='w215' p='IN'>in</w>
    <w id='w216' p='DT'>these</w>
    <w id='w217' p='NNS' l='structure'>structures</w>
    <w id='w218' p='.'>.</w>
   </s>
  </p>
 </text>
 <markup>
  <nes>
   <ne id='e75' fr='w211' to='w211' t='Substance' st='Individual_protein'>
    <textspan>Beta-catenin</textspan>
   </ne>
   <ne id='e77' fr='w217' to='w217' t='Source' st='Cell_component'>
    <textspan>structures</textspan>
   </ne>
  </nes>
  <rels>
   <rel id='r32' e1='e75' e2='e77' t='Causal' st='Change/Location'/>
  </rels>
 </markup>
</doc>
...
</rexml>
```

Figure A.2: *Example document with basic RE XML markup.*

A.2 Conversion to RE XML

The BioInfer data (see Chapter 3) is already encoded in XML, includes sentence and word token markup, and uses token standoff annotation for entities and relations. Therefore, conversion to the RE XML document type is a matter of simple XML-to-XML transformation. Additionally, while the information is not used for the evaluation in this thesis, NOT relations (specifying negation) are mapped to a negation attribute (`neg`) on relation elements (`rel`). And, EQUAL and COREFER relations are converted to coreference information in the form of a grounded entity identifier attribute (`gid`) on entity elements (`ne`).

The ACE data (see Chapter 3) is encoded in SGML, does not include sentence or word token markup, and uses character standoff annotation for entities and relations. Therefore, conversion to RE XML requires SGML-to-XML conversion, tokenisation and mapping from character offset to token offset. Sentence boundary identification and word tokenisation is performed using LT-TTT (Grover et al., 2000), a general purpose text tokenisation tool based on the generic XML text processing tools in LT-XML2 (Grover et al., 2006). Then, the conversion from character to token standoff is performed using LT-XML2 tools. This is achieved by first wrapping each character with an element and then mapping the standoff from the character elements to the word token elements. After this, the data is well-formed XML using token offsets and the final conversion is a simple XML-to-XML transformation. Entity start (`fr`) and end (`to`) tokens are set based on the head extent from the ACE markup.

A.3 Encoding Dependency Parse Information

Figure A.3 contains the extended document type definition for marking up dependency parse information. The top-level element for dependency parse information is `dps`, which is added as a daughter of the `markup` element in the basic RE document type definition in Figure A.1 above. The `dps` element is a container element used to group the individual dependency parse elements (`dp`) corresponding to sentences (`s`) in the document text. The `dp` contain `dpg` elements corresponding to dependency relations where the `d` attribute specifies the dependent word token element and the `g` attribute specifies the governing word token element. The type of the governor-dependency relation is encoded in the `t` attribute and the word token is encoded in the `w` attribute.

Figure A.4 illustrates the Minipar (see Chapter 3) dependency parse for the ex-

ample sentence. Word tokens constitute nodes in the dependency graph, arks specify relations where the word token at the end of the arrow is the dependent token and the annotations (e.g., s+obj) between arrow heads and word tokens specify the relation types. Figure A.5 contains the RE XML markup for corresponding to the dependency parse. The first `dpg` element, for example, encodes a relation of type *object* (s+obj) between the dependent token (`d`) with identifier "w211" and the governor token (`g`) with identifier "w214".

```
<!ELEMENT dps (dp)>          <!-- Dps: Dependency Parse Container          -->
<!ELEMENT dp (dpg*)>         <!-- Dp: Contains Dependency Parse            -->
<!ELEMENT dpg EMPTY>         <!-- Dpg: Specs Governor-Dependency Relation -->

<!ATTLIST dp sid CDATA #REQUIRED>          <!-- DP Sentence ID             -->

<!ATTLIST dpg d CDATA #REQUIRED>           <!-- DPG Dependency Token ID    -->
<!ATTLIST dpg g CDATA #REQUIRED>           <!-- DPG Governor Token ID      -->
<!ATTLIST dpg t CDATA #REQUIRED>           <!-- DPG Gov-Dep Relation Type  -->
<!ATTLIST dpg w CDATA #REQUIRED>           <!-- DPG Word Text              -->
```

Figure A.3: *Additional document type information for encoding dependency parse information.*

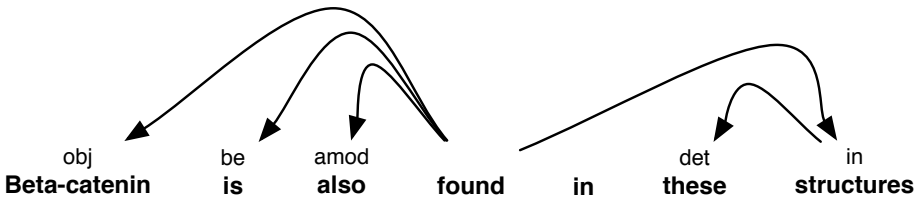


Figure A.4: *Example dependency parse.*

```
<dp sid='s1'>
 <dpg d='w211' g='w214' t='s+obj' w='Beta-catenin'/>
 <dpg d='w212' g='w214' t='be' w='is'/>
 <dpg d='w213' g='w214' t='amod' w='also'/>
 <dpg d='w214' g='' t='' w='found'/>
 <dpg d='w215' g='' t='' w='in'/>
 <dpg d='w216' g='w217' t='det' w='these'/>
 <dpg d='w217' g='w214' t='in' w='structures'/>
 <dpg d='w218' g='' t='' w='.'/>
</dp>
```

Figure A.5: *RE XML markup example dependency parse.*

Appendix B

Full Relation Schemas for Data Sets

B.1 ACE 2004

This section contains details the full relation type schema for the ACE 2004 data. For further details, refer to the annotation guidelines (LDC, 2004c). The total number of mentions for each type is shown in parentheses.

1. PHYSICAL (1216)
   Physical relations describe physical proximities of taggable entities.

   - LOCATED (745)
     The Located relation captures the exact location of an entity. However, if an entity is located in a geographical region like a lake, a river, or a mountain, it should be reported as a Located relation even if the text does not explicitly refer to the shores of the lake, the banks of the river, or the foothills of the mountain.
   - NEAR (87)
     Near indicates that an entity is explicitly near another entity, but not actually in that location or part of that location.
   - PART-WHOLE (384)
     Part-Whole characterizes physical relationships between entities and their parts.

2. PERSONAL-SOCIAL (365)
   Personal-Social relations describe the relationship between entities of type person. No other entity type is allowed as an argument of these relations. The order of the arguments does not impact relations of this type. We record only that there exists a relationship between the entities.

   - BUSINESS (179)
     Business captures the connection between two entities in any professional relationship. This includes boss-employee, lawyer-client, co-workers, political relationships, etc.

- FAMILY (130)
  Family captures the relation between an entity and another entity with which it is in any familial position.
- OTHER (56)
  Other is reserved for all Social relationships that do not cleanly fit into the subtypes above.

3. EMPLOYMENT-MEMBERSHIP-SUBSIDIARY (1631)
   This relation includes: 1) Employment relationship between a person and the organisation or GPE by which they are employed (only valid when the person is paid by the organisation or GPE); 2) Subsidiary relationships (i.e., ownership, administrative, and other hierarchical relationships) between two organizations and between an organization and a GPE; and 3) Membership relationships between an agent (person, organisation, GPE) and an organization of which they are a member.

   - EMPLOY-EXEC(S) (503)
     This subtype describes relations between persons and organizations where the person holds a managerial position such as CEO , president, vice-president, director, leader, or head.
   - EMPLOY-STAFF (554)
     This subtype is for relationships between organizations and GPEs and persons who fill general staff positions within them.
   - EMPLOY-UNDETERMINED (79)
     At times the context does not give you enough in formation to determine whether an individual is performing a managerial or general staff position within an organization. Employ-undetermined is for these relations.
   - MEMBER-OF-GROUP (192)
     Member relations include organization membership such as political party membership, church membership, and so on.
   - SUBSIDIARY (209)
     Subsidiary characterizes the relationship between a company and its parent company.
   - PARTNER (12)
     Partner characterizes the collaborative relationship between two agents (person, organisation, GPE).
   - OTHER (82)
     Other is reserved for relationships between person, organisation, and GPE that do not fit into the other schemas.

4. AGENT-ARTIFACT (212)
   Agent-Artifact describes the relationship between agentive entities and artifacts.

   - USER-OWNER (200)
     An agent is in a Possessor-Owner relationship with an artifact when that agent is the owner of the artifact or has possession of or habitually uses it. In the following

example, it is not explicitly clear whether I own or rent the house. Possessor-Owner can be applied to either relationship.

- INVENTOR-MANUFACTURER (9)
  An agent is in an Inventor-Manufacturer relationship with an artifact when that agent caused the artifact to come into being.
- OTHER (3)
  Other is reserved for any Agent-Artifact relations that do not fall under the other two subtypes.

5. PERSON-ORGANISATION AFFILIATION (142)
   Person-Organisation Affiliation describes relationships between entities that are not captured by other relation types.

   - ETHNIC (39)
     Ethnic describes the relationship between Person(s) and the collective person group to which they are identified.
   - IDEOLOGY (49)
     Ideology describes the relationship between Person(s) and the collective Person/ Organisation group(s) defined by coherent ideological systems to which they are identified by themselves or the article.
   - OTHER (54)
     Other should be used for all Person/Organisation Affiliation relations that do not fit cleanly into any other categories. Many of the relations that fall under this subtype will be cases where a person or organisation modifies another entity. The intended meaning of this construction is often unclear. This subtype can also be filled with relations that have type overlap.

6. GPE AFFILIATION (529)
   GPE Affiliation describes the relationship between entities of type person and organisation with GPEs when more than one aspect of the GPE is referenced by the context of the text.

   - CITIZEN-RESIDENT (273)
     Citizen-Resident describes the relation between a person and the GPE in which they have citizenship or in which they live.
   - BASED-IN (216)
     Organizations are not always located in the GPE in which they are based. We distinguish between the physical locations of an organisation with their GPE of origin with the Based-In Subtype.
   - OTHER (40)
     Other should be used for all GPE Affiliation relations that do not fit cleanly into any other categories. Many of the relations that fall under this subtype will be cases where a GPE modifies another entity. The intended meaning of this construction is often unclear.

7. DISCOURSE (279)
   A Discourse relation is one where a semantic part-whole or membership relation is established only for the purposes of the discourse. The whole or group referred to is not an official entity relevant to world knowledge. Instead, it has been constructed for solely the purposes of discursive efficiency.

B.2 ACE 2005

This section contains details the full relation type schema for the ACE 2005 data. For further details, refer to the annotation guidelines (LDC, 2005b). The total number of mentions for each type is shown in parentheses.

1. PHYSICAL (878)
   - LOCATED (774)
     The Located Relation captures the physical location of an entity. This Relation is restricted to people. In other words, arg1 in Located Relations can only be occupied by mentions of Entities of Type Person.
   - NEAR (104)
     Near indicates that an entity is explicitly near another entity, but neither entity is a part of the other or located in/at the other.
2. PART-WHOLE (643)
   - GEOGRAPHICAL (446)
     The Geographical Relation captures the location of a Facility, Location, or GPE in or at or as a part of another Facility, Location, or GPE. Geographical relationships are the sorts of things one might find in a gazetteer, on a map, or on a building plan (although this is not a requirement per se). Similarly, these are typically permanent relationships, though there are obviously exceptions (a tent might be put up in a certain location for a special event, for example).
   - SUBSIDIARY (184)
     Subsidiary captures the ownership, administrative, and other hierarchical relationships between organizations and between organizations and GPEs. This includes relationships between a company and its parent company, as well as between a department of an organization and that organization. It also includes the relationship between organizations and the GPE's government of which they are a part.
   - ARTIFACT (13)
     Artifact characterizes physical relationships between concrete physical objects and their parts. Both arguments must have the same entity type (though not subtype). This Relation is restricted to Vehicles, Substances, and Weapons.
3. PERSONAL-SOCIAL (360)
   Personal-Social relations describe the relationship between people. Both arguments must be entities of type person. Please note: The arguments of these Relations are not ordered. The Relations are symmetric.

- BUSINESS (77)
  The Business Relation captures the connection between two entities in any professional relationship. This includes boss-employee, lawyer-client, studentteacher, co-workers, political relationships on a personal level, etc. This does not include relationships implied from interaction between two entities (e.g., "President Clinton met with Yasser Arafat last week").
- FAMILY (244)
  The Family Relation captures the connection between one entity and another with which it is in any familial relationship.
- LASTING-PERSONAL (39)
  Lasting-Personal captures relationships that meet the following conditions: 1) The relationship must involve personal contact (or a reasonable assumption thereof) and 2) There must be some indication or expectation that the relationship exists outside of a particular cited interaction. The first condition excludes relationships like "Gore's supporters," "her opponents," or "people who help Americans laugh," where there is no expectation that one party will have interacted personally with the other party (or, put another way, spent time with the other party). A reasonable expectation of personal interaction is sufficient: there are relationships that often but not always involve personal contact (like "classmate" or "neighbor") – these will be allowed in general, as long as their commonplace usage would tend to imply personal contact. The second condition excludes relationships like "his visitors," "his victims," or "his successor," where there is no indication from the text that the relationship exists outside of the specific event being discussed (a visit, a crime, or a succession, here). In the same way, this excludes cases where one might try to infer a relationship from a description of an event involving both entities (e.g., "He visited her in the hospital."). Relationships must be explicitly mentioned in the text.

4. ORG-AFFILIATION (1023)

- EMPLOYMENT (761)
  Employment captures the relationship between Persons and their employe rs. This Relation is only taggable when it can be reasonably assumed that the person is paid by the organisation or GPE. This Relation includes the relationship between an elected representative and the GPE he represents, for example, "John Kerry (D-Massachusetts)."
- OWNERSHIP (15)
  Ownership captures the relationship between a Person and an Organization owned by that Person.
- FOUNDER (31)
  Founder captures the relationship between an agent (Person, Organization, or GPE) and an Organization or GPE established or set up by that agent.
- STUDENT-ALUM (18)
  Student-Alum captures the relationship between a Person and an educational institution the Person attends or attended. Please note that only attendance is required. It is not necessary for the person to have officially graduated from the institution.

- SPORTS-AFFILIATION (24)
  Sports-Affiliation captures the relationship between a player, coach, manager, or assistant and his or her affiliation with a sports organization (including sports leagues or divisions as well as individual sports teams). This Relation subtype exists because it often requires domain-specific world knowledge to determine whether a sports team is made up of paid or unpaid players (i.e. whether a relationship between a player and a team qualifies as Employment).
- INVESTOR-SHAREHOLDER (25)
  Investor-Shareholder captures the relationship between an agent (Person, Organization, or GPE) and an Organization in which the agent has in vested or in which the agent owns shares/stock. Please note that agent s may invest in GPEs.
- MEMBERSHIP (149)
  Membership captures the relationship between an agent and an organization of which the agent is a member. Organizations and GPEs can be members of other Organizations (such as NATO or the UN). As discussed above, instances where a Person is a member of an elected government body (the Senate, the Knesset, the Supreme Court, etc.) will be tagged as Membership, even when the word "member" is not present (e.g., Supreme Court justice).

5. AGENT-ARTIFACT (358)

   - USER-OWNER-INVENTOR-MANUFACTURER (358)
     This Relation applies when an agent owns an artifact, has possession of an artifact, uses an artifact, or caused an artifact to come into being.

6. GEN-AFFILIATION (396)

   - CITIZEN-RESIDENT-RELIGION-ETHNICITY (258)
     Citizen-Resident-Religion-Ethnicity describes the Relation between a person entity and a) the GPE in which they have citizenship, b) the GPE or Location in which they live, c) the religious organisation or person entity with which they have affiliation, or d) the GPE or person entity that indicates their ethnicity. We consider a person's birthplace as a place of residence for this purpose (e.g., "the Russian-born athlete" or "he was born in San Francisco").
   - ORG-LOCATION-ORIGIN (138)
     Org-Location-Origin captures the relationship between an organization and the LOC or GPE where it is located, based, or does business.

B.3 BioInfer

This section contains details of the full relation type schema for the BioInfer data. For further details, refer to the paper describing the corpus (Pyysalo et al., 2007) and the project web page.[1] The total number of mentions for each type is shown in parentheses.

[1] http://mars.cs.utu.fi/BioInfer/?q=relationship_ontology

1. OBSERVATION (145)
   - TEMPORAL (12)
     - COOCCUR (8)
       Use as COOCCUR(*arg1,arg2,...,argN*). The arguments are events that occur together.
     - COEXPRESS (4)
       Use as COEXPRESS(*arg1,arg2,...,argN*). The arguments (genes) are expressed together.
   - SPATIAL (105)
     - COPRECIPITATE (5)
       Use as COPRECIPITATE(*arg1,arg2,...,argN*) where the arguments are proteins. The arguments precipitate as a complex.
     - PRESENCE (8)
       Use as PRESENCE(*arg1,arg2*). An observation that arg1 is present when arg2 occurs. Experimental setups or the presence in a cell are not included.
     - COLOCALIZE (89)
       Use as COLOCALIZE(*arg1,arg2,...,argN*). The arguments (proteins) are found in the same place (or move to the same place) at the same time.
     - ABSENCE (3)
       Use as ABSENCE(*arg1,arg2*). An observation that arg1 is absent when arg2 occurs. Experimental setups or the absence in a cell are not included.
   - COREGULATE (3)
     Use as COREGULATE(*arg1,arg2,...,argN*). The arguments are coregulated.
   - CORELATE (25)
     Use as CORELATE(*arg1,arg2,...,argN*). A general, unspecified co-relation between the arguments.
2. PART-OF (575)
   - COLLECTION:MEMBER (258)
     - MEMBER (258)
       Use as MEMBER(*arg1,arg2*). A member (arg2) belongs to a collection (arg1). For example a protein belongs to a protein family.
   - OBJECT:COMPONENT (317)
     - SUBSTRUCTURE (14)
       Use as SUBSTRUCTURE(*arg1,arg2*). A component (*arg2*) is a part of a structure other than a complex (*arg1*). For example a polymer contains many monomers.
     - F-CONTAIN (13)
       Use as F-CONTAIN(*arg1,arg2*). Like CONTAIN but arg1 is a fusion protein.
     - MUTUALCOMPLEX (10)
       Use as MUTUALCOMPLEX(*arg1,arg2,...,argN*). Like BIND but the arguments may form several complexes, each complex having a different composition.

  - CONTAIN (280)
    Use as CONTAIN(*arg1,arg2*). A component (*arg2*) is a part of a complex (*arg1*).

3. IS-A (517)

   - SIMILARITY (517)
     - FUNCTIONAL-SIMILARITY (15)
       * FNSIMILAR (15)
         Use as FNSIMILAR(*arg1,arg2,...,argN*). A functional similarity. The arguments (proteins) have similar functions. Use only if functional similarity cannot be expressed through interactions with other entities.
     - PHYSICAL-SIMILARITY (75)
       * STSIMILAR (5)
         Use as STSIMILAR(*arg1,arg2,...,argN*). A structural similarity. The arguments (proteins) have similar structures. For example two proteins have a domain in common.
       * SQSIMILAR (14)
         Use as SQSIMILAR(*arg1,arg2,...,argN*). A sequence similarity. The arguments (genes/proteins) have similar sequences.
       * SIMILAR (56)
         Use as SIMILAR(*arg1,arg2,...,argN*). A general, unspecified similarity between the arguments.
     - EQUALITY (427)
       * ENCODE (33)
         Use as ENCODE(*arg1,arg2*). A gene (*arg1*) produces an mRNA or a protein (*arg2*).
       * EQUAL (249)
         Use as EQUAL(*arg1,arg2*). Only for aliases.
       * COREFER (145)
         Use as COREFER(*arg1,arg2*). An anaphoric equality where *arg1* is the anaphora and *arg2* the referent.

4. CAUSAL (1461)

   - CONDITION (117)
     - PREVENT (18)
       Use as PREVENT(*arg1,arg2*). The *arg1* prevents *arg2* from happening. The condition for using this predicate is that *arg2* must not have been happening before. See also FULL-STOP.
     - CONDITION (97)
       Use as CONDITION(*arg1,arg2*). The *arg1* is required (but is not necessarily sufficient) for the *arg2*.
     - MUTUALCONDITION (2)
       Use as MUTUALCONDITION(*arg1,arg2*). Consider as non-directional CONDITION. MUTUALCONDITION(a,b) translates to CONDITION(a,b) CONDITION(b,a).

- CHANGE (1135)
  - DYNAMICS (312)
    - START (24)
      - INITIATE (24)
        Use as INITIATE(*arg1,arg2*) where *arg2* is a process. This predicate is used when *arg2* has not been happening and is now started by *arg1*. See also HALT and STIMULATE.
    - NEGATIVE (101)
      - DOWNREGULATE (15)
        Use as DOWNREGULATE(*arg1,arg2*) where arg2 is a gene (or more specifically gene expression). The arg1 decreases the expression level (i.e. the rate at which the product is produced) of a gene (arg2). See also UPREGULATE and REGULATE.
      - SUPPRESS (70)
        Use as SUPPRESS(*arg1,arg2*) where *arg2* is a process. The *arg1* decreases the speed of the process (*arg2*). See also STIMULATE and CONTROL.
      - INHIBIT (16)
        Use as INHIBIT(*arg1,arg2*) where *arg2* is a protein. The *arg1* decreases the activity (e.g., enzymatic activity) of the protein (*arg2*). See also ACTIVATE, MODULATE, and INACTIVATE.
    - UNSPECIFIED (68)
      - REGULATE (1)
        Use as REGULATE(*arg1,arg2*). A general regulatory relationship where *arg2* is a gene expression. See also UPREGULATE and DOWNREGULATE.
      - CONTROL (57)
        Use as CONTROL(*arg1,arg2*). A general regulatory relationship where *arg2* is a process. See also STIMULATE and SUPPRESS.
      - MODULATE (10)
        Use as MODULATE(*arg1,arg2*). A general regulatory relationship where *arg2* is a protein. See also ACTIVATE and INHIBIT.
    - POSITIVE (109)
      - CATALYZE (4)
        Use as CATALYZE(*arg1,arg2*). The *arg1* (an enzyme) catalyzes *arg2* (a reaction).
      - UPREGULATE (6)
        Use as UPREGULATE(*arg1,arg2*) where *arg2* is a gene (or more specifically gene expression). The *arg1* increases the expression level (i.e. the rate at which the product is produced) of the gene (*arg2*). See also DOWNREGULATE and REGULATE.
      - STIMULATE (41)
        Use as STIMULATE(*arg1,arg2*) where *arg2* is a process. The *arg1* increases the speed of the process (*arg2*). See also SUPPRESS and CONTROL.

- ◇ MEDIATE (41)
  Use as MEDIATE(*arg1,arg2*). The *arg1* mediates *arg2* but does not necessarily regulate it.
- ◇ ACTIVATE (17)
  Use as ACTIVATE(*arg1,arg2*) where *arg2* is a protein. The *arg1* increases the activity (e.g., enzymatic activity) of a protein (*arg2*). See also INHIBIT and MODULATE.

* FULL-STOP (10)
  - ◇ HALT (9)
    Use as HALT(*arg1,arg2*) where *arg2* is a process. This predicate is used when *arg2* has been happening and is now stopped by *arg1*. See also INITIATE and SUPPRESS.
  - ◇ INACTIVATE (1)
    Use as INACTIVATE(*arg1,arg2*) where *arg2* is a protein or a gene. This predicate is used when a gene expression or protein activity (*arg2*) is decreased to essentially zero by *arg1*. This can be considered as an extremely strong downregulation or inhibition. See also DOWNREGULATE and INHIBIT.

– PHYSICAL (508)

* MODIFICATION (40)
  - ◇ REMOVAL (4)
    - · REMOVE (0)
      Use as REMOVE(*arg1,arg2*) where *arg2* is a protein. A general relationship in which *arg1* modifies *arg2* by removing a (small) molecule from it. See also ADD.
    - · DEPHOSPHORYLATE (4)
      Use as DEPHOSPHORYLATE(*arg1,arg2*) where *arg2* is a protein. A phosphate group is removed from *arg2* by *arg1*. See also PHOSPHORYLATE.
  - ◇ ADDITION (30)
    - · ACETYLATE (6)
      Use as ACETYLATE(*arg1,arg2*) where *arg2* is a protein. An acetyl group is added to *arg2* by *arg1*.
    - · ADD (1)
      Use as ADD(*arg1,arg2*) where *arg2* is a protein. A general relationship in which *arg1* modifies *arg2* by adding a (small) molecule to it. See also REMOVE.
    - · PHOSPHORYLATE (23)
      Use as PHOSPHORYLATE(*arg1,arg2*) where *arg2* is a protein. A phosphate group is added to *arg2* by *arg1*. See also DEPHOSPHORYLATE.
  - ◇ MODIFY (6)
    Use as MODIFY(*arg1,arg2*). An unspecified modification where *arg1* modifies *arg2*.

* BREAK-DOWN (28)
  - UNBIND (9)
    Use as UNBIND(*arg1,arg2*). *arg1* and *arg2* dissociate from each other. See also BIND.
  - CLEAVE (4)
    Use as CLEAVE(*arg1,arg2*) where *arg2* is a a protein or a gene. *arg2* is physically cleaved into two by *arg1*.
  - DISASSEMBLE (4)
    Use as DISASSEMBLE(*arg1,arg2*). A general relationship for describing catabolic reactions where *arg1* degrades *arg2*. See also ASSEMBLE.
  - DEPOLYMERIZE (6)
    Use as DEPOLYMERIZE(*arg1,arg2*). Components (monomers) of *arg2* (a polymer) are cleaved/removed from *arg2* by *arg1*. See also POLYMERIZE.
  - DISRUPT (5)
    Use as DISRUPT(*arg1,arg2*) where *arg2* is a complex. This predicate is used when *arg1* makes *arg2* to dissociate to its components. See also BIND.
* ASSEMBLY (440)
  - ATTACH (6)
    Use as ATTACH(*arg1,arg2,...,argN*). A general relationship for describing anabolic reactions where the arguments join together to form a new structure.. See also ASSEMBLE.
  - ASSEMBLE (0)
    Use as ASSEMBLE(*arg1,arg2*). A general relationship for describing anabolic reactions where *arg1* synthesises *arg2*. See also DISASSEMBLE.
  - CROSS-LINK-AP (6)
    Use as CROSS-LINK-AP(*arg1,arg2*). The *arg1* causes the *arg2* to cross-link.
  - CROSS-LINK (9)
    Use as CROSS-LINK(*arg1,arg2*). *arg1* and *arg2* are proteins and are covalently bound. See also BIND.
  - BIND (416)
    Use as BIND(*arg1,arg2,...,argN*). Non-covalent binding (i.e. formation of a complex, association) between the arguments. Each argument is present in the same complex but there are not necessarily all pairwise contacts between the arguments. See also CROSS-LINK and DISRUPT.
  - POLYMERIZE (3)
    Use as POLYMERIZE(*arg1,arg2*). The *arg1* joins multiple *arg2* (monomers) covalently together to form a chain (a polymer). See also DEPOLYMERIZE.

- AMOUNT (10)
  - DECREASE (4)
    Use as DECREASE(*arg1,arg2*). The absolute amount of *arg2* is decreased by *arg1*. See also INCREASE.
  - INCREASE (6)
    Use as INCREASE(*arg1,arg2*). The absolute amount of *arg2* is increased by *arg1*. See also DECREASE.
- LOCATION (68)
  - LOCALIZE (39)
    Use as LOCALIZE(*arg1,arg2*). This predicate is used when *arg1* causes the position of *arg2* to change. See also LOCALIZE-TO.
  - LOCALIZE-TO (29)
    Use as LOCALIZE-TO(*arg1,arg2*). This predicate is used when *arg1* changes its position to *arg2*. See also LOCALIZE.
- AFFECT (117)
  Use as AFFECT(*arg1,arg2*). A general directional relationship where *arg1* causes a change in *arg2*. See also INTERACT.
- INTERACT (119)
  Use as INTERACT(*arg1,arg2,...,argN*). A general non-directional relationship where each argument causes a change in the other arguments . See also AFFECT.
- MUTUAL-AFFECT (1)
  Use as MUTUAL-AFFECT(*arg1,arg2*). Consider as non-directional AFFECT. MUTUAL-AFFECT(a,b) translates to AFFECT(a,b) AFFECT(b,a).

- PARTICIPATE (97)
  Use as PARTICIPATE(*arg1,arg2*). The *arg1* is involved in the *arg2* but is not alone sufficient to cause it.
- CAUSE (101)
  Use as CAUSE(*arg1,arg2*). A general directional causal relationship. The *arg1* is the cause for the *arg2*.
- XOR (11)
  Use as XOR(*arg1,arg2*). The arguments are mutually exclusive.

5. HUMANMADE (10)
   Use as HUMANMADE(*arg1,arg2*). This predicate is used when the relationship is forced or caused by human intervention. The actual type of the relationship is not stated but is one of the types in this ontology.

6. NOT (163)
   Use as NOT(*arg1*) where *arg1* is a relationship. This predicate is used when a relationship is explicitly stated not to be true.

7. RELATE (99)
   Use as RELATE(*arg1,arg2*). A general, unspecified, non-directional relationship. This predicate is used when no details of the relationship is known.

8. REL-ENT (50)
   Use as REL-ENT(*arg1,arg2*). This predicate is used when an unnamed entity (*arg1*) refers to that of a named entity (*arg2*). E.g. REL-ENT("activation","sphingomyelinase") means that "activation" is "activation of sphingomyelinase".

Bibliography

Agichtein, E. (2006). Confidence estimation methods for partially supervised relation extraction. In *Proceedings of the 6th SIAM International Conference on Data Mining*, Bethesda, MD, USA.

Agichtein, E., Cucerzan, S., and Brill, E. (2005). Analysis of factoid questions for effective relation extraction. In *Proceedings of the 28th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Salvador, Brazil.

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

Alex, B., Grover, C., Haddow, B., Kabadjov, M., Klein, E., Matthews, M., Roebuck, S., Tobin, R., and Wang, X. (2008a). Assisted curation: Does text mining really help? In *Proceedings of the Pacific Symposium on Biocomputing*, Hawai'i, HI, USA.

Alex, B., Grover, C., Haddow, B., Kabadjov, M., Klein, E., Matthews, M., Tobin, R., and Wang, X. (2008b). The ITI TXM corpora: Tissue expressions and protein-protein interactions. In *Proceedings of the LREC Workshop on Building and Evaluating Resources for Biomedical Text Mining*, Marrakech, Morocco.

Alex, B., Haddow, B., and Grover, C. (2007). Recognising nested named entities in biomedical text. In *Proceedings of ACL Workshop on Biomedical Natural Language Processing*, Prague, Czech Republic.

Alias-i (2007). LingPipe home. Accessed 30 July 2007 from `http://www.alias-i.com/lingpipe/index.html`.

Anderson, E., Bai, Z., Bischof, C., Blackford, S., Demmel, J., Dongarra, J., Croz, J. D., Greenbaum, A., Hammarling, S., McKenney, A., and Sorensen, D. (1999). *LAPACK Users' Guide*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, third edition. Accessed 25 September 2006 from `http://www.netlib.org/lapack/lug/lapack_lug.html`.

ANSI: American National Standards Institute (1997). *Guidelines for Abstracts*. NISO Press, Bethesda, MD, USA. ANSI/NISO Z39.14-1997.

Artiles, J., Gonzalo, J., and Sekine, S. (2007). The SemEval-2007 WePS evaluation: Establishing a benchmark for the web people search task. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, Prague, Czech Republic.

Banko, M. and Etzioni, O. (2008). The tradeoffs between traditional and open relation extraction. In *Proceedings of 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Columbus, Ohio.

Barzilay, R. and Elhadad, M. (1997). Using lexical chains for text summarization. In *Proceedings of the ACL/EACL Workshop on Intelligent Scalable Text Summarization*, Madrid, Spain.

Barzilay, R., McKeown, K. R., and Elhadad, M. (1999). Information fusion in the context of multi-document summarization. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*, College Park, MD, USA.

Basu, S., Banerjee, A., and Mooney, R. J. (2004). Active semi-supervision for pairwise constrained clustering. In *Proceedings of the SIAM International Conference on Data Mining*, Lake Buena Vista, FL, USA.

Baxendale, P. F. (1958). Man-made index for technical literature: An experiment. *IBM Journal of Research and Development*, 2(5):354–361.

Berry, M. W., Dumais, S. T., and O'Brien, G. W. (1994). Using linear algebra for intelligent information retrieval. *Society for Industrial and Applied Mathematics Review*, 37(4):573–595.

Bienstock, D. and Iyengar, G. (2004). Faster approximation algorithms for packing and covering problems. Technical Report TR-2004-09, Columbia University.

Bikel, D., Schwartz, R., and Weischedel, R. (1999). An algorithm that learns what's in a name. *Machine Learning Journal, Special Issue on Natural Language Engineering*, 34:211–231.

Bilenko, M., Basu, S., and Mooney, R. J. (2004). Integrating constraints and metric learning in semi-supervised clustering. In *Proceedings of the 21st International Conference on Machine Learning*, Banff, Alberta, Canada.

Blaschke, C., Andrade, M. A., Ouzounis, C., and Valencia, A. (1999). Automatic extraction of biological information from scientific text: Protein-protein interactions. In *Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology*, Heidelberg, Germany.

Blei, D., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Blei, D. M., Griffiths, T. L., Jordan, M. I., and Tenenbaum, J. B. (2004). Hierarchical topic models and the nested chinese restaurant process. In Thrun, S., Saul, L. K., and Schölkopf, B., editors, *Proceedings of the 17th Annual Conference on Neural Information Processing Systems*, volume 16. MIT Press, Cambridge, MA.

Blitzer, J., Dredze, M., and Pereira, F. (2006). Domain adaptation with structural correspondence learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia.

Blum, A. and Chawla, S. (2001). Learning from labeled and unlabeled data using graph mincuts. In *Proceedings of the 18th International Conference on Machine Learning*, Williamstown, MA, USA.

Blum, A. and Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computation Learning Theory*, Madison, WI, USA.

Bod, R., Scha, R., and Sima'an, K., editors (2003). *Data-Oriented Parsing*. CSLI Publications, University of Chicago Press, Stanford, CA.

Boguraev, B. and Kennedy, C. (1997). Salience-based content characterization of text documents. In *Proceedings of the ACL/EACL Workshop on Intelligent Scalable Text Summarization*, Madrid, Spain.

Borko, H. and Bernier, C. L. (1975). *Abstracting concepts and methods*. Academic Press, New York.

Bos, J. (2005). Towards wide-coverage semantic interpretation. In *Proceedings of the 6th International Workshop on Computational Semantics*, Tilburg, The Netherlands.

Brandow, R., Mitze, K., and Rau, L. F. (1995). Automatic condensation of electronic publications by sentence selection. *Information Processing and Management*, 31(5):675–685.

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

Briscoe, T. and Carroll, J. (2006). Evaluating the accuracy of an unlexicalized statistical parser on the PARC DepBank. In *Proceedings of the COLING/ACL Main Conference Poster Sessions*, Sydney, Australia.

Briscoe, T., Carroll, J., and Watson, R. (2006). The second release of the RASP system. In *Proceedings of the COLING/ACL Interactive Presentation Sessions*, Sydney, Australia.

Bunescu, R., Ge, R., Kate, R. J., Marcotte, E. M., Mooney, R. J., Ramani, A. K., and Wong, Y. W. (2004). Comparative experiments on learning information extractors for proteins and their interactions. *Artificial Intelligence in Medicine*, 33(2):139–155.

Bunescu, R. C. and Mooney, R. J. (2007). Extracting relations from text: From word sequences to dependency paths. In Kao, A. and Poteet, S., editors, *Natural Language Processing and Text Mining*, pages 29–44. Springer, London.

Carbonell, J. and Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia.

Carreras, X. and Màrquez, L. (2005). Introduction to the ConLL-2005 shared task. In *Proceedings of the 9th Conference on Natural Language Learning*, Ann Arbor, MI, USA.

Charniak, E. (1999). A maximum-entropy-inspired parser. In *Proceedings of the 1st conference on North American chapter of the Association for Computational Linguistics*, Seattle, WA, USA.

Chen, J., Ji, D., Tan, C. L., and Niu, Z. (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.

Chen, J., Ji, D., Tan, C. L., and Niu, Z. (2006). Unsupervised relation disambiguation with order identification capabilities. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia.

Chinchor, N. (1998). Overview of MUC-7. In *Proceedings of the 7th Message Understanding Conference*, Fairfax, VA, USA.

Chu, M., Li, C., Peng, H., and Chang, E. (2002). Domain adaptation for TTS systems. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Orlando, FL, USA.

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

Cohn, D. A., Ghahramani, Z., and Jordan, M. I. (1996). Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145.

Collins, M. (1999). *Head-Driven Statistical Models for Natural Language Parsing*. PhD thesis, University of Pennsylvania.

Collins, M. and Singer, Y. (1999). Unsupervised models for named entity classification. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, College Park, MD, USA.

Conrad, J. G. and Utt, M. H. (1994). A system for discovering relationships by feature extraction from text databases. In *Proceedings of the 17th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia.

Coolican, H. (2004). *Research Methods and Statistics in Psychology*. Hodder & Stoughton, London, fourth edition.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2001). *Introduction to Algorithms*. MIT Press, Cambridge, MA, second edition.

Corston-Oliver, S. (2001). Text compaction for display on very small screens. In *Proceedings of the NAACL Workshop on Automatic Summarization*, Pittsburgh, PA, USA.

Curran, J. R. and Clark, S. (2003). Language independent NER using a maximum entropy tagger. In *Proceedings of the 7th Conference on Natural Language Learning*, Edmonton, Alberta, Canada.

Dagan, I., Lee, L., and Pereira, F. (1997). Similarity-based methods for word sense disambiguation. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, Madrid, Spain.

Daraselia, N., Yuryev, A., Egorov, S., Novichkova, S., Nikitin, A., and Mazo, I. (2004). Extracting human protein interactions from MEDLINE using a full-sentence parser. *Bioinformatics*, 20(5):604–611.

Daumé III, H. (2006). *Practical Structured Learning Techniques for Natural Language Processing*. PhD thesis, University of Southern California.

Daumé III, H. (2007). Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, Prague, Czech Republic.

Daumé III, H. and Marcu, D. (2005). Bayesian multi-document summarization at MSE. In *Proceedings of the ACL Workshop on Multilingual Summarization Evaluation*, Ann Arbor, MI, USA.

de Marneffe, M.-C., MacCartney, B., and Manning, C. D. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, Genoa, Italy.

DeJong, G. (1982). An overview of the FRUMP system. In Lehnert, W. G. and Ringle, M. H., editors, *Strategies for Natural Language Processing*, pages 149–176. Lawrence Erlbaum Associates, Hillsdale, NJ.

Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., and Weischedel, R. (2004). The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Lisbon, Portugal.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Dyer, F. H. (1960). *A compendium of the War of the Rebellion*. T. Yoseloff, New York.

Eckart, C. and Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218.

Edmundson, H. P. (1968). New methods in automatic extracting. *Journal of the ACM*, 16(2):264–285.

Elhadad, N. and McKeown, K. R. (2001). Towards generating patient specific summaries of medical articles. In *Proceedings of NAACL Workshop on Automatic Summarization*, Pittsburgh, PA, USA.

Endres-Niggemeyer, B. (1998). *Summarizing Information*. Springer, Berlin.

Erkan, G. and Radev, D. R. (2004). LexPageRank: Prestige in multi-document text summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain.

Fellbaum, C., editor (1998). *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Filatova, E. and Hatzivassiloglou, V. (2003). Marking atomic events in sets of related texts. In Nicolov, N., Bontcheva, K., Angelova, G., and Mitkov, R., editors, *Recent Advances in Natural Language Processing III*. John Benjamins, Amsterdam/Philadelphia.

Filatova, E. and Hatzivassiloglou, V. (2004). Event-based extractive summarization. In *Proceedings of the ACL Text Summarization Branches Out Workshop*, Barcelona, Spain.

Filatova, E., Hatzivassiloglou, V., and McKeown, K. (2006). Automatic creation of domain templates. In *Proceedings of the COLING/ACL Main Conference Poster Sessions*, Sydney, Australia.

Fundel, K., Küffner, R., and Zimmer, R. (2007). RelEx – relation extraction using dependency parse trees. *Bioinformatics*, 23(3):365–371.

Fung, G. P. C., Yu, J. X., and Lu, H. (2002). Discriminative category matching: Efficient text classification for huge document collections. In *Proceedings of the IEEE International Conference on Data Mining*, Maebashi City, Japan.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.

Ginter, F., Pyysalo, S., Björne, J., Heimonen, J., and Salakoski, T. (2007). BioInfer relationship annotation manual. Technical Report 806, Turku Centre for Computer Science.

Girju, R., Nakov, P., Nastase, V., Szpakowicz, S., Turney, P., and Yuret, D. (2007). SemEval-2007 task 04: Classification of semantic relations between nominals. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, Prague, Czech Republic.

Goldwater, S. and Griffiths, T. L. (2007). A fully Bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, Prague, Czech Republic.

Greenwood, M. and Stevenson, M. (2007). A task-based comparison of information extraction pattern models. In *Proceedings of the ACL Workshop on Deep Linguistic Processing*, Prague, Czech Republic.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101:5228–5235.

Grover, C., Matheson, C., Mikheev, A., and Moens, M. (2000). LT TTT—a flexible tokenisation tool. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, Athens, Greece.

Grover, C., Matthews, M., and Tobin, R. (2006). Tools to address the interdependence between tokenisation and standoff annotation. In *Proceedings of the EACL Workshop on Multi-dimensional Markup in Natural Language Processing*, Trento, Italy.

Hachey, B. (2006). Comparison of similarity models for the relation discovery task. In *Proceedings of the ACL Workshop Linguistic Distances*, Sydney, Australia.

Hachey, B. (2007). Domain-neutral relation characterisation: Evaluation on disease-treatment data. In *Proceedings of ISMB BioLINK SIG Meeting: Linking Literature, Information and Knowledge for Biology*, Vienna, Austria.

Hachey, B., Alex, B., and Becker, M. (2005). Investigating the effects of selective sampling on the annotation task. In *Proceedings of the 9th Conference on Computational Natural Language Learning*, Ann Arbor, MI, USA.

Hachey, B., Grover, C., and Tobin, R. (2008). Datasets for comparative evaluation of relation extraction in the news and biomedical domains. In *Proceedings of Finding the Hidden Knowledge: Text Mining for Biology and Medicine*, Glasgow, Scotland.

Haghighi, A. and Klein, D. (2006). Prototype-driven learning for sequence models. In *Proceedings of the Joint Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics Annual Meeting*, New York, NY, USA.

Haghighi, A. and Klein, D. (2007). Unsupervised coreference resolution in a nonparametric Bayesian model. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic.

Hahn, U. and Reimer, U. (1999). Knowledge-based text summarization: Salience and generalization operators for knowledge base abstraction. In Mani, I. and Maybury, M. T., editors, *Advances in Automatic Text Summarization*, pages 215–232. MIT Press, Cambridge, MA.

Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(2/3):107–145.

Hara, T., Miyao, Y., and Tsujii, J. (2005). Adapting a probabilistic disambiguation model of an HPSG parser to a new domain. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing*, Jeju Island, Korea.

Harabagiu, S., Bejan, C. A., and Morărescu, P. (2005). Shallow semantics for relation extraction. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, Edinburgh, Scotland.

Harabagiu, S. M. and Maiorano, S. J. (2002). Multi-document summarization with GISTexter. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, Las Palmas, Spain.

Harman, D. (1992). The DARPA TIPSTER project. *ACM SIGIR Forum*, 26(2):26–28.

Harman, D. and Marcu, D., editors (2001). *Proceedings of the ACM SIGIR Workshop on Text Summarization*, New Orleans, LA, USA.

Hasegawa, T., Sekine, S., and Grishman, R. (2004). Discovering relations among named entities from large corpora. In *Proceedings of the 42nd Annual Meeting of Association of Computational Linguistics*, Barcelona, Spain.

Hasegawa, T., Sekine, S., and Grishman, R. (2005). Unsupervised paraphrase acquisition via relation discovery. Technical report, Proteus Project, Computer Science Department, New York University.

Hassan, H., Hassan, A., and Noeman, S. (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.

Hatzivassiloglou, V. and McKeown, K. R. (1993). Towards the automatic identification of adjectival scales: Clustering adjectives according to meaning. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, OH, USA.

Hirschman, L., Yeh, A., Blaschke, C., and Valencia, A. (2004). Overview of BioCreAtIvE: Critical assessment of information extraction for biology. In *Proceedings of Critical Assessment of Information Extraction Systems in Biology Workshop (BioCreAtIvE)*, Granada, Spain.

Hochbaum, D. S. (1997). Approximating covering and packing problems: set cover, vertex cover, independent set and related problems. In Hochbaum, D. S., editor, *Approximation Algorithms for NP-Hard Problems*, pages 94–143. PWS Publishing Company, Boston, MA.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42:177–196.

Jenssen, T.-K., Lægreid, A., Komorowski, J., and Hovig, E. (2001). A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*, 28(1):21–28.

Jing, H., Kambhatla, N., and Roukos, S. (2007). Extracting social networks and biographical facts from conversational speech transcripts. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic.

Johnson, M. (2007). Why doesn't EM find good HMM POS-taggers? In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Prague, Czech Republic.

Jones, R., Ghani, R., Mitchell, T., and Riloff, E. (2003). Active learning with multiple view feature sets. In *Proceedings of the ECML Workshop on Adaptive Text Extraction and Mining*, Cavtat-Dubrovnik, Croatia.

Jurafsky, D. and Martin, J. H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice-Hall, Upper Saddle River, NJ.

Karamanis, N. (2007). Integrating natural language processing with FlyBase curation. In *Proceedings of the Pacific Symposium on Biocomputing*, Maui, HI, USA.

Klein, D. (2005). *The Unsupervised Learning of Natural Language Structure*. PhD thesis, Stanford University.

Klein, D., Kamvar, S., and Manning, C. (2002). From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering. In *Proceedings of the 19th International Conference on Machine Learning*, Sydney, Australia.

Knowles, J. H. J. and Kell, D. B. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics*, 21(15):3201–3212.

Kupiec, J., Pedersen, J., and Chen, F. (1995). A trainable document summarizer. In *Proceedings of the 18th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Seattle, WA, USA.

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25:259–284.

Lange, T., Braun, M. L., Roth, V., and Buhmann, J. M. (2003). Stability-based model selection. In Becker, S., Thrun, S., and Obermayer, K., editors, *Advances in neural information processing systems*, volume 15, pages 617–624. MIT Press, Cambridge, MA.

LDC (2004a). *The ACE 2005 (ACE05) Evaluation Plan*. Linguistic Data Consortium. Accessed 23 July 2008 from `http://www.nist.gov/speech/tests/ace/ace05/doc/ace05-evalplan.v3.pdf`.

LDC (2004b). *Annotation Guidelines for Entity Detection and Tracking (EDT)*. Linguistic Data Consortium. Accessed 22 July 2008 from `http://www.ldc.upenn.edu/Projects/ACE/docs/EnglishEDTV4-2-6.PDF`.

LDC (2004c). *Annotation Guidelines for Relation Detection and Characterization (RDC)*. Linguistic Data Consortium. Accessed 22 July 2008 from `http://www.ldc.upenn.edu/Projects/ACE/docs/EnglishRDCV4-3-2.PDF`.

LDC (2005a). *ACE (Automatic Content Extraction) English Annotation Guidelines for Entities*. Linguistic Data Consortium. Accessed 22 July 2008 from http://www.ldc.upenn.edu/Projects/ACE/docs/English-Entities-Guidelines_v5.6.1.pdf.

LDC (2005b). *ACE (Automatic Content Extraction) English Annotation Guidelines for Relations*. Linguistic Data Consortium. Accessed 22 July 2008 from http://www.ldc.upenn.edu/Projects/ACE/docs/English-Relations-Guidelines_v5.8.3.pdf.

Lee, L. (1999). Measures of distributional similarity. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, College Park, MD, USA.

Leskovec, J., Milic-Frayling, N., and Grobelnik, M. (2005). Impact of linguistic analysis on the semantic graph coverage and learning of document extracts. In *Proceedings of the 20th National Conference On Artificial Intelligence*, Pittsburgh, PA, USA.

Li, W., Wu, M., Lu, Q., Xu, W., and Yuan, C. (2006). Extractive summarization using inter- and intra- event relevance. In *Proceedings of the Joint 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia.

Lin, C.-Y. (2004). ROUGE: a package for automatic evaluation of summaries. In *Proceedings of the ACL Text Summarization Branches Out Workshop*, Barcelona, Spain.

Lin, C.-Y. and Hovy, E. (2000). The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th conference on Computational linguistics*, Saarbrücken, Germany.

Lin, C.-Y. and Hovy, E. (2003). Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the Joint Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics Annual Meeting*, Edmonton, Alberta, Canada.

Lin, D. (1997). Using syntactic dependency as local context to resolve word sense ambiguity. In *Proceedings of the 35th Annual Meeting on Association for Computational Linguistics*, Madrid, Spain.

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

Lin, D. and Pantel, P. (2001). Discovery of inference rules for question answering. *Natural Language Engineering*, 7(4):343–360.

Liu, Y., Jin, R., Jain, A., and Mallapragada, P. (2007a). BoostCluster: Boosting clustering by pairwise constraints. In *Proceedings of the 13th International Conference on Knowledge Discovery and Data Mining*, San Jose, CA, USA.

Liu, Y., Shi, Z., and Sarkar, A. (2007b). Exploiting rich syntactic information for relationship extraction from biomedical articles. In *Proceedings of the Joint Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics Annual Meeting*, Rochester, NY, USA.

Lombardi, M., Hobbs, R., Richards, J., and Hobbs, R. C. (2003). *Mark Lombardi: Global Networks*. Independent Curators Inc., New York, NY.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2).

Luo, X. (2005). On coreference resolution performance metrics. In *Proceedings of the Joint Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Vancouver, BC, Canada.

Mani, I. (2001). *Automatic Summarization*. John Benjamins, Amsterdam/Philadelphia.

Manning, C. D., Raghavan, P., and Schtze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Marcu, D. (1997). From discourse structures to text summaries. In *Proceedings of the ACL/EACL Workshop On Intelligent Scalable Text Summarization*, Madrid, Spain.

Marcu, D. (2006). Automatic discourse parsing. In Brown, K., editor, *Encyclopedia of Language & Linguistics*, pages 683–696. Elsevier, Boston.

Matuszek, C., Cabral, J., Witbrock, M., and DeOliveira, J. (2006). An introduction to the syntax and content of Cyc. In *Proceedings of the 2006 AAAI Spring Symposium on Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answereing*, Stanford, CA, USA.

McCallum, A., Corrada-Emmanuel, A., and Wang, X. (2004). The author-recipient-topic model for topic and role discovery in social networks: Experiments with Enron and academic email. Technical report, University of Massachusetts Amherst.

McDonald, R. (2007). A study of global inference algorithms in multi-document summarization. In *Proceedings of the 29th European Conference on Information Retrieval*, Rome, Italy.

McKeown, K. and Radev, D. R. (1995). Generating summaries of multiple news articles. In *Proceedings of the 18th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Seattle, WA, USA.

McKeown, K. R., Jordan, D. A., and Hatzivassiloglou, V. (1998). Generating patient-specific summaries of online literature. In *Proceedings of the AAAI Spring Symposium on Intelligent Text Summarization*, Stanford, CA, USA.

Mel'čuk, I. (1987). *Dependency Syntax: Theory and Practice*. State University of New York Press, Albany, NY.

Mihalcea, R. (2005). Language independent extractive summarization. In *Proceedings of the ACL Poster/Demonstration Sessions*, Ann Arbor, MI, USA.

Miike, S., Itoh, E., Ono, K., and Sumita, K. (1994). A full-text retrieval system with a dynamic abstract generation function. In *Proceedings of the 17th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Dublin, Ireland.

Minnen, G., Carroll, J., and Pearce, D. (2000). Robust, applied morphological generation. In *Proceedings of the 1st International Natural Language Generation Conference*, Mitzpe Ramon, Israel.

MUC-5 (1993). *Proceedings of the 5th Message Understanding Conference*, Baltimore, MD, USA. Morgan Kaufmann, San Mateo, CA.

MUC-6 (1995). *Proceedings of the 6th Message Understanding Conference*, Columbia, MD, USA. Morgan Kaufmann, San Mateo, CA.

MUC-7 (1998). *Proceedings of the 7th Message Understanding Conference*, Fairfax, VA, USA. Accessed 19 July 2008 from `http://www-nlpir.nist.gov/related_projects/muc/proceedings/muc_7_toc.html`.

Murray, G., Renals, S., and Carletta, J. (2005a). Extractive summarization of meeting recordings. In *Proceedings of the 9th European Conference on Speech Communication and Technology*, Lisbon, Portugal.

Murray, G., Renals, S., Carletta, J., and Moore, J. (2005b). Evaluating automatic summaries of meeting recordings. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, Ann Arbor, MI, USA.

Nakao, Y. (2000). An algorithm for one-page summarization of a long text based on thematic hierarchy detection. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, Hong Kong.

Ng, A., Jordan, M., and Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. In Dietterich, T. G., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press, Cambridge, MA.

NIST (2006). NIST 2005 automatic content extraction evaluation official results (ACE05). Accessed 11 July 2007 from `http://www.nist.gov/speech/tests/ace/ace05/doc/ace05eval_official_results_20060110.htm`.

Niu, Z.-Y., Ji, D.-H., and Tan, C.-L. (2004). Document clustering based on cluster validation. In *Proceedings of the 13th ACM International Conference on Information and Knowledge Management*, Washington, DC, USA.

Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S., and Yuret, D. (2007). The CoNLL 2007 shared task on dependency parsing. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Prague, Czech Republic.

Nobata, C., Sekine, S., Isahara, H., and Grishman, R. (2002). Summarization system integrated with named entity tagging and IE pattern discovery. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, Las Palmas, Spain.

Peirce, C. S. (1870). Description of a notation for the logic of relatives, resulting from an amplification of the conceptions of Boole's calculus of logic. *Memoirs of the American Academy of Sciences*, 9:317–378.

Peirce, C. S. (1933). The simplest mathematics. In Hartshorne, C. and Weiss, P., editors, *Collected Papers of Charles Sanders Peirce*, volume 4. Harvard University Press, Cambridge, MA.

Popescu-Belis, A. and Robba, I. (1998). Three new methods for evaluating reference resolution. In *Proceedings of the LREC Workshop on Linguistic Coreference*, Madrid, Spain.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program: Electronic Library and Information Systems*, 14(3):130–137.

Pustejovsky, J., Castaño, J., Ingria, R., Saurí, R., Gaizauskas, R., Setzer, A., and Katz, G. (2003). TimeML: Robust specification of event and temporal expressions in text. In *Proceedings of the 5th International Workshop on Computational Semantics*, Tilburg, The Netherlands.

Pyysalo, S., Ginter, F., Heimonen, J., Björne, J., Boberg, J., Järvinen, J., and Salakoski, T. (2007). BioInfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8(50).

Pyysalo, S., Ginter, F., Pahikkala, T., Boberg, J., Järvinen, J., and Salakoski, T. (2006). Evaluation of two dependency parsers on biomedical corpus. *International Journal of Medical Informatics, special issue on Recent Advances in Natural Language Processing for Biomedical Applications*, 75(6):430–442.

Pyysalo, S., Ginter, F., Pahikkala, T., Koivula, J., Boberg, J., Järvinen, J., and Salakoski, T. (2004). Analysis of link grammar on biomedical dependency corpus targeted at protein-protein interactions. In *Proceedings of the COLING International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, Geneva, Switzerland.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Montr'eal, Qu'ebec, Canada.

Riedel, S. and Klein, E. (2005). Genic interaction extraction with semantic and syntactic chains. In *Proceedings of the 4th Learning Language in Logic Workshop*, Bonn, Germany.

Riloff, E. (1996). Automatically generating extraction patterns from untagged text. In *Proceedings of the 14th National Conference on Artificial Intelligence*, Portland, OR, USA.

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

Rindflesch, T. C., Tanabe, L., Weinstein, J. N., and Hunter, L. (2000). EDGAR: extraction of drugs, genes and relations from the biomedical literature. In *Proceedings of the Pacific Symposium on Biocomputing*, Oahu, HI, USA.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smyth, P. (2004). The author-topic model for authors and documents. In *20th Conference on Uncertainty in Artificial Intelligence*, Banff, Alberta, Canada.

Rush, J. E., Salvador, R., and Zamora, A. (1971). Automatic abstracting and indexing. II. Production of indicative abstracts by application of contextual inference and syntactic coherence criteria. *Journal of the American Society for Information Science*, 22(4):260–274.

Rutter, J. D. (1994). A serial implementation of Cuppen's divide and conquer algorithm for the symmetric eigenvalue problem. Technical Report CSD-94-799, Computer Science Division (EECS), University of California, Berkely.

Saggion, H. and Lapalme, G. (2002). Generating indicative-informative summaries with SumUM. *Computational Linguistics*, 28(4):497–526.

Salton, G. and McGill, M. J. (1986). *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, NY.

Sang, E. F. T. K. and Meulder, F. D. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the 7th Conference on Natural Language Learning*, Edmonton, Alberta, Canada.

Sanguinetti, G., Laidler, J., and Lawrence, N. D. (2005). Automatic determination of the number of clusters using spectral algorithms. In *Proceedigns of the IEEE Workshop on Machine Learning for Signal Processing*, Mystic, CT, USA.

Schulte im Walde, S. (2003). *Experiments on the Automatic Induction of German Semantic Verb Classes*. PhD thesis, Universität Stuttgart.

Schwartz, A. S. and Hearst, M. A. (2003). A simple algorithm for identifying abbreviation definitions in biomedical text. In *Proceedings of the Pacific Symposium on Biocomputing*, Lihue, HI, USA.

Sekine, S. (2001). *OAK System - Manual*. Department of Computer Science, New York University. Accessed 24 July 2008 from `http://nlp.cs.nyu.edu/oak/manual.html`.

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

Seung, H. S., Opper, M., and Sompolinsky, H. (1992). Query by committee. In *Proceedings of the 5th Annual Workshop on Computational Learning Theory*, Pittsburgh, PA, USA.

Smalheiser, N. R. and Swanson, D. R. (1998). Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. *Computer Methods and Programs in Biomedicine*, 57(3):149–153.

Smith, D. A. (2002). Detecting and browsing events in unstructured text. In *Proceedings of the 25th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Tampere, Finland.

Smith, N. A. (2006). *Novel Estimation Methods for Unsupervised Discovery of Latent Structure in Natural Language Text*. PhD thesis, Johns Hopkins University.

Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.

Spärck Jones, K. (1999). Automatic summarising: Factors and directions. In Mani, I. and Maybury, M. T., editors, *Advances in Automatic Text Summarisation*, pages 1–14. MIT Press, Cambridge, MA.

Spärck Jones, K. (2007). Automatic summarising: The state of the art. *Information Processing and Management*, 43:1449–1481.

Sparck Jones, K. and Galliers, J. R. (1996). *Evaluating Natural Language Processing Systems: An Analysis and Review*. Springer-Verlag, Secaucus, NY, USA.

Stapley, B. J. and Benoit, G. (2000). Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts. In *Proceedings of the Pacific Symposium on Biocomputing*, Oahu, HI, USA.

Steinberger, J., Kabadjov, M. A., Poesio, M., and Sanchez-Graillet, O. (2005). Improving LSA-based summarization with anaphora resolution. In *Proceedings of the Joint Human Language Technology Conference and Conference on Empirical Methods in Natural Language*, Vancouver, BC, Canada.

Steyvers, M. and Griffiths, T. L. (2007). Probabilistic topic models. In Landauer, T. K., McNamara, D. S., Dennis, S., and Kintsch, W., editors, *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum Associates, Mahwah, NJ, USA.

Strehl, A. and Ghosh, J. (2003). Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617.

Sudo, K., Sekine, S., and Grishman, R. (2003). An improved extraction pattern representation model for automatic IE pattern acquisition. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, Sapporo, Japan.

Swanson, D. R. (1986). Fish oil, raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, 30:7–18.

Tan, P.-N., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining*. Addison Wesley, Boston, MA, USA, first edition.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2004). Hierarchical Dirichlet processes. Technical report, University of California, Berkeley.

Teufel, S. and Moens, M. (1997). Sentence extraction as a classification task. In *Proceedings of the ACL/EACL Workshop on Intelligent and Scalable Text Summarization*, Madrid, Spain.

Teufel, S. and Moens, M. (2002). Summarizing scientific articles – experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445.

Thompson, C. A., Califf, M. E., and Mooney, R. J. (1999). Active learning for natural language parsing and information extraction. In *Proceedings of the 16th International Conference on Machine Learning*, Bled, Slovenia.

Tomita, J., Soderland, S., and Etzioni, O. (2006). Expanding the recall of relation extraction by bootstrapping. In *Proceedings of the EACL Workshop on Adaptive Text Extraction and Mining*, Trento, Italy.

Torii, M., Liu, H., Hu, Z., and Wu, C. (2006). A comparison study of biomedical short form definition detection algorithms. In *Proceedings of the CIKM Workshop on Text Mining in Bioinformatics*, Arlington, VA, USA.

Trouilleux, F., Gaussier, Éric., Bès, G. G., and Zaenen, A. (2000). Coreference resolution evaluation based on descriptive specificity. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, Athens, Greece.

Tucker, R. I. and Spärck Jones, K. (2005). Between shallow and deep: An experiment in automatic summarising. Technical Report 632, Computer Laboratory, University of Cambridge.

Turmo, J., Ageno, A., and Català, N. (2006). Adaptive information extraction. *ACM Computing Surveys*, 38(2).

Turney, P. D. (2006). Similarity of semantic relations. *Computational Linguistics*, 32(3):379–416.

van Valin, Jr., R. D. (2006). Functional relations. In Brown, K., editor, *Encyclopedia of Language & Linguistics*, pages 683–696. Elsevier, Boston.

Vanderwende, L., Banko, M., and Menezes, A. (2004). Event-centric summary generation. In *Proceedings of the Document Understanding Workshop*, Boston, MA, United States.

Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Katz, G., and Pustejovsky, J. (2007). SemEval-2007 task 15: TempEval temporal relation identification. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, Prague, Czech Republic.

Vlachos, A. and Gasperin, C. (2006). Bootstrapping and evaluating named entity recognition in the biomedical domain. In *Proceedings of the HLT-NAACL Workshop on Linking Natural Language Processing and Biology: Towards Deeper Biological Literature Analysis*, New York, NY, USA.

von Luxburg, U. (2006). A tutorial on spectral clustering. Technical Report 149, Max Planck Institute for Biological Cybernetics.

Waibel, A., Bett, M., and Finke, M. (1998). Meeting browser: Tracking and summarising meetings. In *Proceedings of the DARPA Broadcast News Transcription And Understanding Workshop*, Lansdowne, VA, USA.

Wellner, B., McCallum, A., Peng, F., and Hay, M. (2004). An integrated, conditional model of information extraction and coreference with application to citation matching. In *20th Conference on Uncertainty in Artificial Intelligence*, Banff, Alberta, Canada.

White, M. and Cardie, C. (2002). Selecting sentences for multidocument summaries using randomized local search. In *Proceedings of the ACL Workshop on Automatic Summarization*, Philadelphia, PA, USA.

White, M., Korelsky, T., Cardie, C., Ng, V., Pierce, D., and Wagstaff, K. (2001). Multidocument summarization via information extraction. In *Proceedings of the 1st International Conference on Human Language Technology Research*, San Diego, CA, USA.

Xing, E. P., Ng, A. Y., Jordan, M. I., and Russell, S. (2003). Distance metric learning, with application to clustering with side-information. In Becker, S., Thrun, S., and Obermayer, K., editors, *Advances in Neural Information Processing Systems*, volume 15. MIT Press, Cambridge, MA.

Yoshioka, M. and Haraguchi, M. (2004). Multiple news articles summarization based on event reference information. In *Proceedings of NTCIR-4 on the Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Summarization*, Tokyo, Japan.

Zelenko, D., Aone, C., and Richardella, A. (2002). Kernel methods for relation extraction. *Journal of Machine Learning Research*, 3:1083–1106.

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

Zhang, M., Su, J., Wang, D., Zhou, G., and Tan, C. L. (2005). Discovering relations from a large raw corpus using tree similarity-based clustering. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing*, Jeju Island, Korea.

Zhang, Z. (2004). Weakly-supervised relation classification for information extraction. In *Proceedings of the 13th International Conference on Information and Knowledge Management*, Washington DC, USA.

Zhang, Z. (2005). Mining inter-entity semantic relations using improved transductive learning. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing*, Jeju Island, Korea.

Zhao, Y. and Karypis, G. (2004). Empirical and theoretical comparisons of selected criterion functions for document clustering. *Machine Learning*, 55:311–331.

Zhao, Y. and Karypis, G. (2005). Hierarchical clustering algorithms for document datasets. *Data Mining and Knowledge Discovery*, 10:141–168.

Zhou, G., Su, J., Zhang, J., and Zhang, M. (2005). Exploring various knowledge in relation extraction. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, Ann Arbor, MI, USA.