Natural Language Engineering **18** (1): 21–59. © Cambridge University Press 2011
doi:10.1017/S1351324911000106

Datasets for generic relation extraction[*]

B. HACHEY[1], C. GROVER[2] and R. TOBIN[2]

[1]*Language Technology Group, Macquarie University, NSW 2109, Australia*
email: bhachey@cmcrc.com
[2]*Informatics Forum, 10 Crichton Street, Edinburgh, EH8 9AB, Scotland*
email: C.Grover@ed.ac.uk; R.Tobin@ed.ac.uk

(*Received 14 January 2010; revised 23 August 2010; accepted 28 January 2011; first published online 9 March 2011*)

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 or RDF triplestore that can be more effectively used for querying and automated reasoning. A number of resources have been developed for training and evaluating automatic systems for relation extraction in different domains. However, comparative evaluation is impeded by the fact that these corpora use different markup formats and notions of what constitutes a relation. We describe the preparation of corpora for comparative evaluation of relation extraction across domains based on the publicly available ACE 2004, ACE 2005 and BioInfer data sets. We present a common document type using token standoff and including detailed linguistic markup, while maintaining all information in the original annotation. The subsequent reannotation process normalises the two data sets so that they comply with a notion of relation that is intuitive, simple and informed by the semantic web. For the ACE data, we describe an automatic process that automatically converts many relations involving nested, nominal entity mentions to relations involving non-nested, named or pronominal entity mentions. For example, the first entity is mapped from 'one' to 'Amidu Berry' in the membership relation described in 'Amidu Berry, one half of PBS'. Moreover, we describe a comparably reannotated version of the BioInfer corpus that flattens nested relations, maps part-whole to part-part relations and maps n-ary to binary relations. Finally, we summarise experiments that compare approaches to generic relation extraction, a knowledge discovery task that uses minimally supervised techniques to achieve maximally portable extractors. These experiments illustrate the utility of the corpora.[1]

1 Introduction

A vast amount of usable electronic data is in the form of unstructured text. The information-extraction (IE) task aims to identify useful information in text and recode it in a format such as a relational database or RDF triplestore that can be more effectively used for querying and automated reasoning (e.g. Turmo, Ageno and

[*] This work was supported by Scottish Enterprise Edinburgh-Stanford Link grant R37588 as part of the EASIE project at the University of Edinburgh.
[1] http://benhachey.info/data/gre/

https://doi.org/10.1017/S1351324911000106 Published online by Cambridge University Press

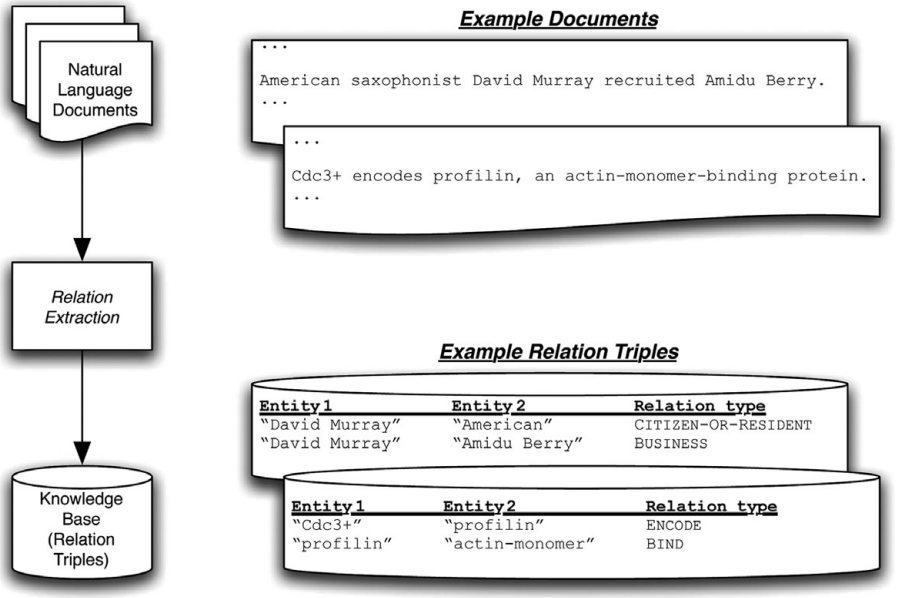


Fig. 1. Overview of relation-extraction task with example input and output.

Català 2006). Typically, information extraction includes the subtasks 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 that is referred to as relation extraction (RE). The RE task has been defined in various ways (e.g. Swanson 1986; Chinchor 1998; Doddington *et al.* 2004; Ginter *et al.* 2007). Here, we aim to identify mentions of relations that are directly expressed in text. A relation mention is defined as a predicate ranging over two arguments, where an argument represents concepts, objects or people and the relation predicate describes the type of association or interaction that holds between the things represented by the arguments. This definition is also informed by the semantic web and the linked data movements that aim to encode knowledge in subject–predicate–object triples that are tractable for large-scale automatic reasoning (e.g. Auer *et al.* 2009; Bizer, Heath and Berners-Lee 2009; Byrne 2009).

Figure 1 contains example relation mentions from the news and biomedical data sets used here. The left side of the figure is a pipeline representation of the RE task. The input consists of natural language documents containing unstructured text. These documents are fed to the RE system, which identifies relations described in the text data and annotates them with a label describing the type of relation. The output of the RE system consists of relation mention tuples, which include the entity mentions that take part in the relation and the relation type. The right side of Figure 1 contains two example input documents on the top and the relation mention tuples from those sentences on the bottom. The first document contains the sentence 'American saxophonist David Murray recruited Amidu Berry'. This

https://doi.org/10.1017/S1351324911000106 Published online by Cambridge University Press

Table 1. *Relation extraction data sets with total number of relations (rels), total number of relation types (types), total number of entity-type pair subdomains (subdomains) and mean number of relation types per subdomain (types per SD)*

Data Set	Dom	Rels	Types	Subs	Mean
Message Understanding Conference (MUC-7)	News	1,612	3	3	1
BioText Disease-Treatment (BTDT)	Med	964	5	1	5
BioText Protein–Protein Interaction (BTPPI)	Bio	1,570	24	1	24
AIMed (AIMed)	Bio	880	1	1	1

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'.

Various corpora have been created for the RE task. Table 1 contains a list of some well-known data sets. The first column (Data set) gives the name of the data set. The second column (Dom) gives the domain. The third column (Rels) gives the total number of relation mentions annotated in the data set. The fourth column (Types) gives the total number of relation types in the annotation schema. The fifth column (Subs) gives the total number of subdomains – entity-type pairs (e.g. PERSON–PERSON, GENE–PROTEIN) for which relations are annotated. Finally, the sixth column (Mean) gives the mean number of relation types per subdomain. Many data sets (e.g. MUC-7, AIMed) have only one relation type per subdomain, so knowing the entity types is a sufficient information to fully specify the relation type. And, while other data sets (e.g. BTDT and BTPPI) have multiple relation types, they have only one subdomain making it impossible to assess reliability across entity-type pairs.

Here, we leverage two corpora that have detailed relation-type schemas, including types that are not determined by entity type. We derive data for the news domain from the corpora prepared for the Automatic Content Extraction (ACE)-shared tasks (Doddington *et al.* 2004). Moreover, we derive data for the biomedical domain from the Bio Information Extraction Resource (BioInfer) corpus (Pyysalo *et al.* 2007). Together, these corpora allow tuning and evaluation of systems addressing the multi-relation RE task, including comparative evaluation across the news and biomedical domains. However, these data sets are in different formats, include different linguistic markup (e.g. BioInfer has sentence markup, while ACE does not) and encode different notions of entities and relations.

The primary contribution of this paper is a series of automatic transformations for deriving versions of ACE and BioInfer that allow comparison of RE results across domains. This standardisation process is summarised in Figure 2. Step 1 takes the raw corpora as input and converts it to a common document type. The result of the refactoring is a simple XML format using token standoff, where character offsets are

https://doi.org/10.1017/S1351324911000106 Published online by Cambridge University Press

(1)	*Refactoring*	Convert to common RE document type.
(2)	*Preprocessing*	Add shallow linguistic information.
(3)	*Preprocessing*	Add dependency parse information.
(4)	*Reannotation*	Normalise to common notion of relation.
(5)	*Example usage*	Prepare data for generic relation extraction.

Fig. 2. Process for standardising RE corpora to enable comparative evaluation.

maintained for full reproducibility of the original annotation (e.g. Grover, Matthews and Tobin 2006). In Step 2, linguistic information is added into the XML format. This includes part-of-speech tags and lemmas, as well as shallow parsing from LT-TTT2.[2] Step 3 adds dependency parse information from Minipar (Lin 1998).

In Step 4, the data is normalised to comply with our common notion of relation that acts as a middle ground between various annotation efforts. We require that relations be binary and between named or pronominal entities where possible. This definition (1) enforces consistency across data sets, (2) allows a principled and tractable definition of the generic relation-extraction (GRE) task addressed in Section 8 and (3) complies with the semantic web and the linked data movements that aim to encode knowledge in subject–predicate–object triples for large-scale automatic reasoning (e.g., Auer *et al.* 2009; Bizer *et al.* 2009; Byrne 2009).

Finally, Step 5 consists of a further standardisation for the GRE experiments in Section 8. GRE is a knowledge discovery task that aims to identify mentions of relations in text using techniques that achieve comparable accuracy when transferred across domains without modification of model parameters. The goal of the GRE data preparation is to produce data sets that are similar in terms of the total number of relation mentions, the number of subsets and the number of relation types per subset.

Note that Steps 1 through 3 are completely general and maintain compatibility with all original annotation while Steps 4 and 5 add alternative entity and relation markup. The distribution will consist of the refactored corpora in the common RE document type from Step 1, including inline shallow linguistic markup from Step 2. The remaining information (i.e. dependency parses from Step 3, normalised relations from Step 4 and GRE relations from Step 5) will be included as external files containing token standoff annotations.

The modified version of the ACE data will be available for redistribution through the Linguistic Data Consortium (LDC) as Edinburgh Regularised ACE (reACE). The respective modified corpora will be available to licence holders for the original distributions for ACE 2004[3] and ACE 2005.[4] The modified version of the BioInfer data will be made available as Edinburgh Regularised BioInfer (reBioInfer) free of charge under the same open-source licence terms as the original BioInfer data set.

[2] `http://www.ltg.ed.ac.uk/software/lt-ttt2`
[3] `http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2005T09`
[4] `http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T06`

https://doi.org/10.1017/S1351324911000106 Published online by Cambridge University Press

Further information and links to downloads can be found at `http://benhachey.info/data/gre/`.

2 Background

2.1 Refactoring and normalising annotated data

Previous efforts have standardised corpora for ease of reuse (Johnson *et al.* 2007; Heimonen *et al.* 2008) and for comparative evaluation (Pyysalo *et al.* 2008). However, they have not investigated corpora as diverse as ACE and BioInfer. Furthermore, they have not required that the data sets have multiple relation types that are not determined by the types of the entities forming a relation.

Johnson *et al.* (2007) report a feasibility study on corpus refactoring. The effort is motivated by a previous study suggesting that format is a primary factor in corpus uptake (Cohen *et al.* 2005). The authors take the Protein Design Group corpus as an example, converting the document-level annotation to character-offset annotation in two popular formats. Our standardisation is also motivated by accessibility. However, normalisation of diverse data sets is an additional key goal, allowing comparison of relation extraction across domains.

Pyysalo *et al.* (2008) describe an effort to normalise four protein–protein interaction corpora to a 'a shared level of information, consisting of undirected, untyped binary interactions.' Like Pyysalo *et al.*, our reannotation of BioInfer flattens nested entities, maps n-ary to binary relation mentions and ignores directionality of relations. Unlike Pyysalo *et al.*, our reannotation additionally maps part-whole to part-part relations. Crucially, where Pyysalo et al. focus exclusively on protein-protein interactions as their only relation type, we have the explicit goal of multi-type relation extraction across domains. Therefore, we maintain all physical entity types and all BioInfer relations over physical entity types. In addition, we consider data from the news and biomedical domains.

Heimonen *et al.* (2008) describe a process for converting the complex, structured annotation of the BioInfer corpus to binary relations. They argue that this is a necessary simplification towards extraction of detailed knowledge about protein–protein interactions and demonstrate that the binarisation is largely valid with limited loss of information. Heimonen *et al.* do not perform other types of normalisations described here. In particular, they do not map part-whole to part-part relations (see Section 6.2). Furthermore, they do not explore compatibility across domains.

2.2 Generic relation extraction (GRE)

Relation extraction can be addressed using supervised, bootstrapping or generic approaches. 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 (e.g. Aone *et al.* 1998; Bunescu *et al.* 2004) incur the highest cost as systems need to be built largely from scratch for each new domain. Bootstrapping approaches (e.g. Brin 1999; Agichtein and Gravano 2000) incur less cost as they require only a small amount of seed data. Finally, generic approaches (e.g. Conrad

https://doi.org/10.1017/S1351324911000106 Published online by Cambridge University Press

Table 2. *Sources for ACE 2004 and ACE 2005 news data*

Source	Type	Epoch	Num of 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%)

and Utt 1994; Hasegawa, Sekine and Grishman 2004) 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, the supervised and bootstrapping approaches address only the ontology instantiation problem, while the 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, generic approaches have high utility in terms of developing cheap components for applications like paraphrase acquisition (Hasegawa, Sekine and Grishman 2005), on-demand information extraction (Sekine 2006) and automatic summarisation (Hachey 2009a).

In Section 8, we present summary experiments for GRE across news and biomedical domains. These experiments illustrate the utility of the standardised corpora described in this paper.

2.3 Data sets

The data for the news domain is derived from the IE corpora that was prepared for the 2004 and 2005 ACE-shared tasks (Mitchell *et al.* 2005; Walker *et al.* 2006).[5] The data discussed here is drawn from the newswire and broadcast news materials prepared for the 2004 and 2005 RE tasks (LDC 2004b, 2005b). Table 2 summarises

[5] `http://projects.ldc.upenn.edu/ace/`

https://doi.org/10.1017/S1351324911000106 Published online by Cambridge University Press

the documents in these corpora. 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. 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 for ACE 2004 and 2005, respectively. In addition, the overall newswire/broadcast news splits are approximately 36.8%/63.2% and 27.2%/72.8%, respectively.

The data for the biomedical domain is derived from the IE corpora, which have been prepared and freely distributed as the BioInfer corpus (Pyysalo *et al.* 2007).[6] This consists of 1,100 sentences that were selected from the PubMed database of biomedical literature.[7] The corpus data was collected by entering known pairs of interacting proteins from the Database of Interacting Proteins (DIP)[8] 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.

3 Refactoring: a common document type for RE data

The first step of our standardisation is to convert the corpora to a common format. Figure 3 contains the RE XML document-type definition developed here. This includes a top-level document (`doc`) element, which is made up of a `text` element and a `markup` element. The `text` element contains the tokenised document text, marked up with inline paragraph (`p`), sentence (`s`) and word token (`w`) information. The `markup` element contains standoff entity (`nes`) and relation (`rels`) annotation. Individual entity mentions (`ne`) are specified with reference to the identifiers of the word tokens that start and end the entity text span (attributes `@fr` and `@to`). When present in the annotation, the head of the entity mention phrase is also specified (attributes `@hfr` and `@hto`). Individual relation instances (`rel`) are specified with reference to the entities participating in the relation (attributes `@e1` and `@e2`).

Entity and relation mention identifiers (`@id`) are the same as those used in the source data and `@gid` attributes specify the id of the resolved set of coreferring entity mentions for a given document. Entity and relation types are specified in `@t` attributes and subtypes, if annotated, in `@st` attributes. All `ne` and `rel` annotations, which do not comply with the RE XML document type, are saved in extra attribute (`exattr`) elements so that no information is lost from the original data set.

[6] `http://mars.cs.utu.fi/BioInfer/`
[7] `http://www.ncbi.nlm.nih.gov/pubmed/`
[8] `http://dip.doe-mbi.ucla.edu/`

https://doi.org/10.1017/S1351324911000106 Published online by Cambridge University Press

```
<!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*,exattr*)>       <!-- Ne: Contains NE Textspan   -->
<!ELEMENT textspan (#PCDATA)>           <!-- Textspan: Contains Text    -->
<!ELEMENT rels (rel*)>                  <!-- Rels: Contains Rel Ment'ns -->
<!ELEMENT rel (textspan?,exattr*)>      <!-- Rel: Contains Rel Textspan -->

<!ELEMENT exattr EMPTY>                 <!-- Exattr: Extra Attribute    -->

<!ATTLIST doc id CDATA #IMPLIED>              <!-- Document ID          -->
<!ATTLIST s id CDATA #REQUIRED>               <!-- Sentence ID          -->
<!ATTLIST w id CDATA #REQUIRED>               <!-- Token ID             -->

<!ATTLIST ne id CDATA #REQUIRED>              <!-- NE Mention ID        -->
<!ATTLIST ne gid CDATA #IMPLIED>              <!-- Grounded NE ID       -->
<!ATTLIST ne fr CDATA #REQUIRED>              <!-- NE Start Token ID    -->
<!ATTLIST ne to CDATA #REQUIRED>              <!-- NE End Token ID      -->
<!ATTLIST ne so CDATA #REQUIRED>              <!-- NE Start Char Offset -->
<!ATTLIST ne eo CDATA #REQUIRED>              <!-- NE End Char Offset   -->
<!ATTLIST ne hfr CDATA #IMPLIED>              <!-- NE Head Start Tok ID -->
<!ATTLIST ne hto CDATA #IMPLIED>              <!-- NE Head End Tok ID   -->
<!ATTLIST ne t CDATA #REQUIRED>               <!-- NE Type              -->
<!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         -->
<!ATTLIST rel neg CDATA #IMPLIED>             <!-- Polarity of relation -->

<!ATTLIST exattr n CDATA #REQUIRED>           <!-- Extra Attr Name      -->
<!ATTLIST exattr v CDATA #REQUIRED>           <!-- Extra Attr Value     -->
```

Fig. 3. Basic document-type definition for RE XML.

3.1 *Refactoring ACE*

The ACE data 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. First, character entity references are converted to standard XML and spaces are added, where necessary, to maintain correct character offsets. ACE does not have token markup. Therefore, we add sentence boundary and word token markup using LT-TTT2,[9] a general purpose text tokenisation tool based on the generic XML text processing tools in LT-XML2 (Grover *et al.* 2006).[10]

Next, 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.

[9] http://www.ltg.ed.ac.uk/software/lt-ttt2
[10] http://www.ltg.ed.ac.uk/software/ltxml2

https://doi.org/10.1017/S1351324911000106 Published online by Cambridge University Press

Occasionally the original ACE markup spans substrings of words, for example, in the possessive 'its', only the substring 'it' is marked up as an entity mention. In cases such as this, the converted markup must point at the tokens rather than token substrings so the converted entity mention is 'its' rather than 'it'. Although we do not utilise it in the experiments here, information about the original character offsets is preserved using the end offset (`@eo`) and start offset (`@so`) attributes on `ne` elements. Thus, the `ne` element for 'its' is marked up with `@eo` '-1'.

After this, the data is well-formed XML using token offsets and the conversion to basic RE XML is a simple XML-to-XML transformation. Figure 4 contains an example document with the basic RE XML markup. This is drawn from the ACE data and contains markup for the following sentence:

'American saxophonist David Murray recruited Amidu Berry and DJ Awadi.'[11]

The markup in the figure specifies five entities (`ne`). The first `ne` element (with `@id` 'e60'), for example, contains the markup for the PERSON entity 'American saxophonist David Murray', which starts (`@fr`) and ends (`@to`), respecitively, with word tokens 'w292' and 'w295' and has 'David Murray' as its head. The markup in the figure also specifies three relations: a `Citizen-or-Resident` relation between 'saxophonist' and 'American', a `Business` relation between 'David Murray' and 'Amidu Berry' and a `Business` relation between 'David Murray' and 'Awadi'. In Section 6 we argue that the `Citizen-or-Resident` relation should actually be between 'David Murray' and 'American' and we describe rules for automatically mapping this and other similar cases where the ACE annotation marks relations with a nominal rather than a named entity mention.

3.2 Refactoring BioInfer

The BioInfer data, 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. In addition, while the information is not used for the experiments here, 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`).

Figure 5 contains an example document with the basic RE XML markup. This is drawn from the BioInfer data and contains markup for the following sentence:

[11] This is a simplified version of the following ACE sentence: 'When American saxophonist David Murray recorded his acclaimed Afrocentric jazz album, Fo Juke Review in Dakar, he recruited Amidu Berry and DJ Awadi from PBS to show what an edge West African music can really have.' The sentence is from the ACE 2004 broadcast news document PRI20001103.2000.2994.sgm, which is a transcript of a Public Radio International broadcast from 3 November 2000.

https://doi.org/10.1017/S1351324911000106 Published online by Cambridge University Press

```
<doc id='15'>
 <text>
  <p>
   <s id='s17'>
    <w id='w292'>American</w>
    <w id='w293'>saxophonist</w>
    <w id='w294'>David</w>
    <w id='w295'>Murray</w>
    <w id='w296'>recruited</w>
    <w id='w297'>Amidu</w>
    <w id='w298'>Berry</w>
    <w id='w299'>and</w>
    <w id='w300'>DJ</w>
    <w id='w301'>Awadi</w>
    <w id='w302'>.</w>
   </s>
  </p>
 </text>
 <markup>
  <nes>
   <ne id='e62' gid='E20' fr='w292' to='w292' t='GPE' st='Nation'>
    <textspan type='extent'>American</textspan>
    <textspan type='head'>American</textspan>
    <exattr n='CLASS' v='SPC'/>
    <exattr n='LDCTYPE' v='PRE'/>
   </ne>
   <ne id='e61' gid='E18' fr='w292' to='w293' hfr='w293' hto='w293' t='PER'>
    <textspan type='extent'>American saxophonist</textspan>
    <textspan type='head'>saxophonist</textspan>
    <exattr n='CLASS' v='SPC'/>
    <exattr n='LDCTYPE' v='PRE'/>
   </ne>
   <ne id='e60' gid='E18' fr='w292' to='w295' hfr='w294' hto='w295' t='PER'>
    <textspan type='extent'>American saxophonist David Murray</textspan>
    <textspan type='head'>David Murray</textspan>
    <exattr n='CLASS' v='SPC'/>
    <exattr n='LDCTYPE' v='NAM'/>
   </ne>
   <ne id='e4' gid='E38' fr='w297' to='w298' t='PER'>
    <textspan type='extent'>Amidu Berry</textspan>
    <textspan type='head'>Amidu Berry</textspan>
    <exattr n='CLASS' v='SPC'/>
    <exattr n='LDCTYPE' v='NAM'/>
   </ne>
   <ne id='e5' gid='E1' fr='w300' to='w301' hfr='w301' hto='w301' t='PER'>
    <textspan type='extent'>DJ Awadi</textspan>
    <textspan type='extent'>Awadi</textspan>
    <exattr n='CLASS' v='SPC'/>
    <exattr n='LDCTYPE' v='NAM'/>
   </ne>
  </nes>
  <rels>
   <rel id='11-1' gid='11' e1='e61' e2='e62' t='GPE-AFF'
        st='Citizen-or-Resident'/>
   <rel id='2-1' gid='2' e1='e60' e2='e4' t='PER-SOC' st='Business'/>
   <rel id='3-1' gid='3' e1='e60' e2='e5' t='PER-SOC' st='Business'/>
  </rels>
 </markup>
</doc>
```

Fig. 4. Example of refactored ACE document with basic RE XML markup.

'Beta-catenin is also found in these structures.'[12]

[12] This is a simplified version of the following BioInfer sentence: 'Accordingly, beta-catenin is also found in these structures, again in the absence of alpha-catenin.' This is sentence 15 in the original BioInfer distribution.

https://doi.org/10.1017/S1351324911000106 Published online by Cambridge University Press

```
<doc id='15'>
 <text>
  <p>
   <s id='s11'>
    <w id='w211'>Beta-catenin</w>
    <w id='w212'>is</w>
    <w id='w213'>also</w>
    <w id='w214'>found</w>
    <w id='w215'>in</w>
    <w id='w216'>these</w>
    <w id='w217'>structures</w>
    <w id='w218'>.</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='e78' fr='w214' to='w215' t='RELATIONSHIP_TEXTBINDING'>
    <textspan>found in</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'>
   </rel>
  </rels>
 </markup>
</doc>
```

Fig. 5. Example of refactored BioInfer document with basic RE XML markup.

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'). Note that not all BioInfer entities consist of continuous token sequences. The phrase 'alpha 5 and beta 1 integrins', for example, is annotated with two entity mentions 'alpha 5 integrins' and 'beta 1 integrins'. For the first mention, the `@fr`, `@to` markup here simply indexes into 'alpha' and 'integrins' and these boundaries are used for the experiments. However, the non-continuous annotation is saved in `exattr` elements.

4 Pre-processing: adding TTT linguistic information

Next, we enrich the data with various types of linguistic pre-processing information. This uses the components available as part of LT-TTT2 to perform part-of-speech (POS) tagging, lemmatisation, identification and interpretation of nominalisations, verb and noun phrase chunking, identification of chunk heads and identification of voice and polarity of verb phrases. The POS tagging component uses the C&C maximum entropy POS tagger (Curran and Clark 2003) trained on data tagged with the Penn Treebank POS tagset (Marcus, Marcinkiewicz and Santorini 1993). The lemmatisation component uses morpha (Minnen, Carroll and Pearce 2000). All other

https://doi.org/10.1017/S1351324911000106 Published online by Cambridge University Press

```
<!ATTLIST w l CDATA #IMPLIED>       <!-- Lemma -->
<!ATTLIST w p CDATA #REQUIRED>      <!-- Part-of-speech -->
<!ATTLIST w phr CDATA #IMPLIED>     <!-- Chunk tag -->
<!ATTLIST w vstem CDATA #IMPLIED>   <!-- Nominalisation stem -->
<!ATTLIST w headv CDATA #IMPLIED>   <!-- Verb group head -->
<!ATTLIST w voice CDATA #IMPLIED>   <!-- Verb group voice -->
<!ATTLIST w neg CDATA #IMPLIED>     <!-- Verb group negation -->
```

Fig. 6. Additional document-type information for encoding dependency parse information.

```
<doc id='15'>
 <text>
  <p>
   <s id='s17'>
    <w l='american' p='NNP' phr='B-NP'>American</w>
    <w l='saxophonist' p='NN' phr='I-NP'>saxophonist</w>
    <w l='david' p='NNP' phr='B-NP'>David</w>
    <w l='murray' p='NNP' phr='I-NP'>Murray</w>
    <w l='recruit' p='VBD' phr='B-VP' voice='act'>recruited</w>
    <w l='amidu' p='NNP' phr='B-NP'>Amidu</w>
    <w l='berry' p='NNP' phr='I-NP'>Berry</w>
    <w p='CC' phr='I-NP'>and</w>
    <w l='dj' p='NNP' phr='I-NP'>DJ</w>
    <w l='awadus' p='NNP' phr='I-NP'>Awadi</w>
    <w p='NN'>.</w>
   </s>
  </p>
 </text>
 ...
</doc>
```

Fig. 7. RE XML markup for TTT linguistic information (ACE example).

LT-TTT2 components are rule-based components implemented using the LT-XML2 tools.

Figure 9 contains the extended document-type definition for marking up shallow linguistic information. The linguistic information from TTT is included as an attributive markup on word elements. Lemmas and parts-of-speech are included, respectively, in @l and @p attributes. Noun and verb phrase information from shallow parsing is included in the @phr attribute. This is represented using standard BI markup, where 'B-*X*' indicates that a word is the beginning of a phrase of type *X* and 'I-*X*' indicates that a word is inside a phrase of type *X*. When a noun is a nominalisation (e.g. 'inspiration'), its verb stem (e.g. 'inspire') is given in the @vstem attribute. The main word in a verb phrase is indicated by a 'yes' value for the w attribute @headv. Verb group voice (i.e. 'active' or 'passive') is included in the @voice attribute on the main word of a verb phrase. Negative polarity is indicated by a 'yes' value for the @neg attribute.

4.1 Adding TTT linguistic information to ACE

Figure 7 contains the RE XML markup corresonding to the linguistic information from TTT for the ACE example sentence. The markup in the figure specifies one verb phrase 'recruited' indicated by the 'B-VP' value of the @phr attribute on the corresponding w element. The markup for 'recruited' also indicates the following: the

https://doi.org/10.1017/S1351324911000106 Published online by Cambridge University Press

```
<doc id='15'>
 <text>
  <p>
   <s id='s11'>
    <w l='beta-catenin' p='NN' phr='B-NP'>Beta-catenin</w>
    <w l='be' p='VBZ' phr='B-VP'>is</w>
    <w p='RB' phr='I-VP'>also</w>
    <w l='find' p='VBN' phr='I-VP' headv='yes' voice='pass'>found</w>
    <w p='IN'>in</w>
    <w p='DT' phr='B-NP'>these</w>
    <w l='structure' p='NNS' phr='I-NP'>structures</w>
    <w p='.'>.</w>
   </s>
  </p>
 </text>
 ...
</doc>
```

Fig. 8. RE XML markup for TTT linguistic information (BioInfer example).

```
<!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 cr CDATA #REQUIRED>          <!-- DPG Collapsed Gov-Dep Rel -->
<!ATTLIST dpg w CDATA #REQUIRED>           <!-- DPG Word Text             -->
<!ATTLIST dpg p CDATA #REQUIRED>           <!-- DPG Word Part-of-speech   -->
<!ATTLIST dpg l CDATA #REQUIRED>           <!-- DPG Word Lemma            -->
```

Fig. 9. Additional document-type information for encoding dependency parse information.

part-of-speech is past tense verb (@p='VBD'), the lemma is 'recruit' (@l='recruit') and the verb phrase is in active voice (@voice='act'). In addition, the markup in the figure specifies three noun phrases 'American saxophonist', 'David Murray' and 'Amidu Berry and DJ Awadi'.

4.2 Adding TTT linguistic information to BioInfer

Figure 8 contains the RE XML markup corresponding to the linguistic information from TTT for the BioInfer example sentence. The markup in the figure specifies one verb phrase 'is also found' with main verb 'found' indicated by the @headv attribute on the corresponding w element. The markup for 'found' also indicates the following: the part-of-speech is past participle verb (@p='VBN'), the lemma is 'find' (@l='find') and the verb phrase is passive (@voice='pass'). In addition, the markup in the figure specifies two noun phrases, 'Beta-catenin' and 'these structures'.

5 Pre-processing: adding dependency parse information

Pre-processing also includes dependency parsing. Figure 9 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

https://doi.org/10.1017/S1351324911000106 Published online by Cambridge University Press

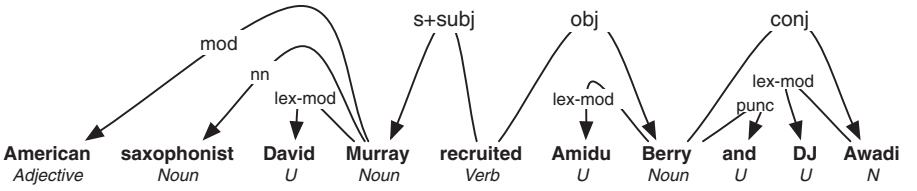


Fig. 10. Dependency parse for an example of ACE sentence.

of the `markup` element in the basic RE document-type definition in Figure 3 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` elements contain `dpg` elements corresponding to individual governor-dependency relations where the `@d` attribute specifies the dependent word token element and the `@cr` attribute specifies the governor. The governor is encoded as `GovType:TokenID`, where `GovType` specifies the type of the governor-dependency relation and `TokenID` specifies the governing word token element (cf. examples in Sections 5.1 and 5.2). Note that the `@cr` attribute encodes collapsed dependency relations that are a product of postprocessing operations over *antecedent* and *preposition complement* relations (described in Subsections 5.1 and 5.2). Finally, the word token is encoded in the `@w` attribute.

For the current work, dependency parses are obtained from the Minipar software.[13] Minipar 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 (Lin 1998). It achieves approximately 79% coverage of the dependency relationships in the SUSANNE corpus with 89% precision. In addition to the directional link from governors to their dependent lexical items and the associated grammatical relation types, Minipar produces parts-of-speech and lemmas, which are stored in the `@p` and `@l` attributes, respectively.

Note that Minipar is used as an illustration here and as a dependency parsing proof-of-concept for the GRE experiments in Section 8. It is chosen because it is accurate, efficient and used widely. Output from other dependency parsers could also be encoded using the representation here. For example, the refactored version of BioInfer includes the Link Grammar markup from the original distribution. Error analysis in related work (Hachey 2009b) addresses the extent to which Minipar performance degrades performance on biomedical RE data.

5.1 Adding dependency parse information to ACE

Figure 10 contains the Minipar dependency parse of the example ACE 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+subj) between arrow heads and word tokens specify the relation types. Dependency relations include, e.g. a *modifier* (mod) relation from governor

[13] `http://webdocs.cs.ualberta.ca/~lindek/minipar.htm`

https://doi.org/10.1017/S1351324911000106 Published online by Cambridge University Press

```
<dp sid='s17'>
 <dpg d='w292' cr='mod:w295' w='American' p='A' l='~'/>
 <dpg d='w293' cr='nn:w295' w='saxophonist' p='N' l='~'/>
 <dpg d='w294' cr='lex-mod:w295' w='David' p='U' l='~'/>
 <dpg d='w295' cr='s+subj:w296' w='Murray' p='N' l='David Murray'/>
 <dpg d='w296' cr='' w='recruited' p='V' l='recruit'/>
 <dpg d='w297' cr='lex-mod:w298' w='Amidu' p='U' l='~'/>
 <dpg d='w298' cr='obj:w296' w='Berry' p='N' l='Amidu Berry'/>
 <dpg d='w299' cr='punc:w298' w='and' p='U' l='~'/>
 <dpg d='w300' cr='lex-mod:w301' w='DJ' p='U' l='~'/>
 <dpg d='w301' cr='conj:w298' w='Awadi' p='N' l='DJ Awadi'/>
 <dpg d='w302' cr='' w='.' p='U' l='~'/>
</dp>
```

Fig. 11. RE XML markup example of ACE dependency parse.

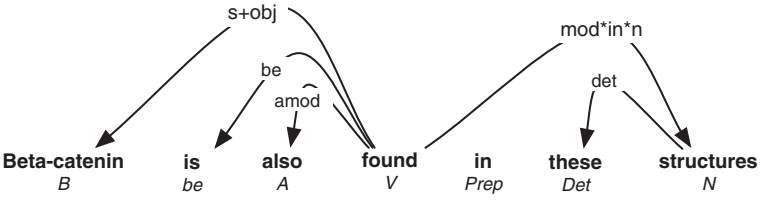


Fig. 12. Dependency parse for example of BioInfer sentence.

noun 'Murray' to dependent adjective 'American', and an *object* (obj) relation from 'recruited' to dependent noun 'Berry'. Note that the *subject* (s+subj) relation from governor verb 'recruited' to dependent noun 'Murray' is not actually a single relation in the Minipar output. It orignially consists of a *surface subject* (s) relation from governor verb 'recruited' to dependent noun 'Murray' and a *subject* (subj) relation from governor verb 'recruited' to an introduced empty node marked as being coreferent with the noun 'Murray'. These are collapsed to a single s+subj relation as a postprocessing step. This serves to connect verbs directly to arguments with deep grammatical relation types. Figure 11 contains the RE XML markup corresponding to the dependency parse. The first `dpg` element, for example, encodes the *modifier* relation between 'American' (@`cr`='mod:w295') and 'Murray' (@`d` 'w292').

5.2 *Adding dependency parse information to BioInfer*

Figure 12 illustrates the Minipar dependency parse of the BioInfer example sentence.[14] Again, 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. Dependency relations include, e.g. a *surface subject*/*object* (s+obj) relation from 'found' to 'Beta-catenin', and a *nominal in modifier* (mod*in*n) from 'found' to 'structures'. Note that the *nominal in modifier* (mod*in*n) relation from governor

[14] While we prefer Minipar for our experiments, the original BioInfer corpus also includes parses from the Link Grammar parser (`http://www.link.cs.cmu.edu/link/`).

https://doi.org/10.1017/S1351324911000106 Published online by Cambridge University Press

```
<dp sid='s11'>
 <dpg d='w211' cr='s+obj:w214' w='Beta-catenin' p='N' l='Beta-catenin'/>
 <dpg d='w212' cr='be:w214' w='is' p='be' l='be'/>
 <dpg d='w213' cr='amod:w214' w='also' p='A' l='~'/>
 <dpg d='w214' cr='' w='found' p='V' l='find'/>
 <dpg d='w215' cr='' w='in' p='Prep' l='~'/>
 <dpg d='w216' cr='det:w217' w='these' p='Det' l='~'/>
 <dpg d='w217' cr='mod*in*n:w214' w='structures' p='N' l='structure'/>
</dp>
```

Fig. 13. RE XML markup example of BioInfer dependency parse.

verb 'found' to dependent noun 'structures' is not actually a single grammatical relation in the Minipar output. It originally consists of two relations: a *modifier* (mod) relation from governor 'found' to dependent 'in' and a *nominal preposition complement* (pcomp-n) relation from governor 'in' to dependent 'structures'. These are collapsed to a single relation as a post-processing step following Lin and Pantel (2000). This serves to connect the prepositional complement directly to the words modified by the preposition. Figure 13 contains the RE XML markup corresponding to the dependency parse. The first `dpg` element, for example, encodes a relation of type *object* (s+obj) between the governor token with identifier 'w214' (`@cr`='s+obj:w214') and the dependent token (`@d`) with identifier 'w211' (i.e. 'Beta-catenin' is the object of 'found').

6 Reannotation: a common notion of relations

The next step is reannotation, where the data sets are normalised to comply with a common notion of relation. Here a relation mention is defined as follows:

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.

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 that are generally more discrete in nature, describing things that happen or occur (e.g. Pustejovsky *et al.* 2004). The choice of binary relations (1) enforces consistency across data sets, (2) allows a principled and tractable definition of the GRE task addressed in the experiments here and (3) complies with the semantic web and the linked data movements, which aim to encode knowledge in subject–predicate–object triples that are tractable for large-scale automatic reasoning (e.g. Auer *et al.* 2009; Bizer *et al.* 2009; Byrne 2009).

Furthermore, we specify that relations should be between named or pronominal entities wherever possible (more in Section 6.1). And, we specify that part-whole and part-part relations should be consistenly marked (more in Section 6.2). Again, this is an intentionally pragmatic definition of IE where the goal is to contribute relation information to scalable online applications of natural language processing and deep semantic interpretation is not strictly necessary.

https://doi.org/10.1017/S1351324911000106 Published online by Cambridge University Press

Table 3. *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*

Label	Description	Example	ACE 2004		ACE 2005	
NAM	Named entity reference	'John', 'Fargo'	6,903	(30.4%)	4,586	(25.4%)
PRO	Pronominal reference	'they', 'her'	5,119	(22.5%)	4,684	(25.9%)
NOM	Nominal reference	'the lawyer'	4,853	(21.3%)	4,001	(22.1%)
PRE	Prenominal reference	'[Labour] nominee'	2,992	(13.2%)	2,489	(13.8%)
BAR	Unquantified nominals	'lawyers'	1,990	(8.8%)	1,673	(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%)

6.1 Reannotating ACE

Reannotation of ACE 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 facilitated by the presence of detailed linguistic annotation in the ACE corpus, which makes it possible to automatically map many nominal entity mentions to named entity mentions. Several aspects of the ACE annotation are used in the mapping rules: entity extent, entity type and entity mention type. 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]]'.

This contains one nominal and two named entity mentions:

(1) 'Amidu Barry' with type Person (PER) and mention type *named* (NAM).
(2) 'one half of PBS' with type Person and mention type *nominal* (NOM).
(3) 'PBS' with type Organisation (ORG) and mention type *named*.

We will also talk about embedding and embedded entity mentions. In the above example, 'one half of PBS' is embedding and 'PBS' is embedded.

Table 3 contains the full list of possible entity mention types with the number and proportion of occurrences in the ACE 2004 and the ACE 2005 data sets. The rules used here only consider nominal and prenominal entity mentions for possible mapping. Unquantified nominal mentions are generally not coreferent with named entity mentions and other mention types are too rare. The ACE 2004 entity types include Person (PER), Organisation (ORG), Facility (FAC), Location (LOC), Geographical/Political (GPE), Vehicle (VEH) (LDC 2004c). The ACE 2005 entity types include Person (PER), Organisation (ORG), Geographical/Social/Political (GPE), Location (LOC), Facility (FAC), Vehicle (VEH) and Weapon (WEA) (LDC 2005a).

https://doi.org/10.1017/S1351324911000106 Published online by Cambridge University Press

Entity phrase heads and coreference are also annotated in ACE. Head markup is based on the notion of headedness in syntactic grammars, where a head is the word or category that gets propagated 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. Heads are indicated here by an underscore, e.g. 'one' in 'one half of PBS'.[15] Coreference markup indicates the entity 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. Finally, some of the mapping rules described below make use of aspects of the linguistic preprocessing introduced in Section 4.

Figure 14 contains three example mappings from different rules. Mapping rule 1 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 relation mention EXECUTIVE ('Commons', 'Commons speaker Michael Martin') 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 4. 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.

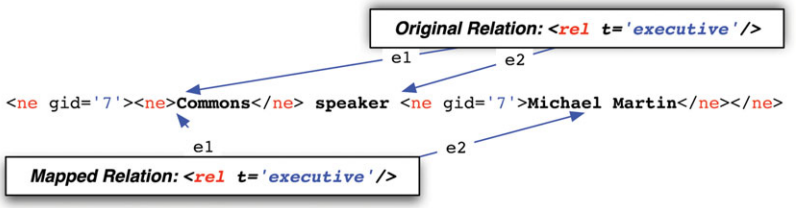
Rule 2 maps to a coreferent and embedded entity mention occurring immediately to the left of the head of the original entity mention, for example:

'$[^{gpl}_{nom}$ $[^{gpl}_{nam}$ Indonesia]'s war-torn $[^{gpl}_{nam}$ Aceh] province]'

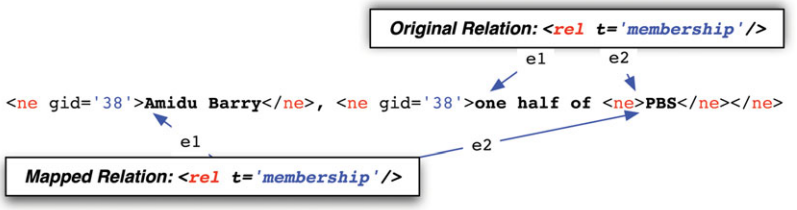
PART-WHOLE('Indonesia's war-torn Aceh province', 'Indonesia')
PART-WHOLE('Aceh', 'Indonesia').

[15] Arguably, the annotator should have marked 'half' or 'one half' as the head of 'one half of PBS'. However, this does not affect the mapping rules described here.

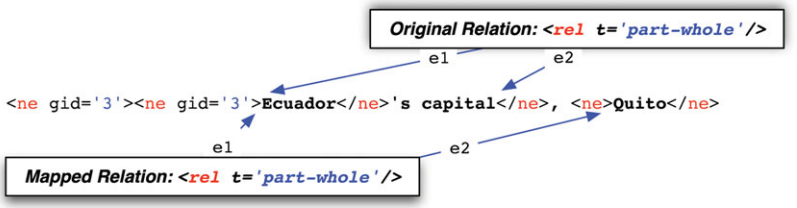
https://doi.org/10.1017/S1351324911000106 Published online by Cambridge University Press



(a) Mapping rule 1: Prenominal ↦ embedded coreferent



(b) Mapping rule 5: Nominal ↦ left adjacent coreferent



(c) Mapping rule 6: Nominal ↦ right adjacent coreferent

Fig. 14. Example of ACE mappings from nominal to named entity mentions.

Rule 3 maps to any coreferent and embedded entity mention, for example:

‘$[^{gpl}_{nom}$ $[^{per}_{nam}$ gore]’s home state of $[^{gpl}_{nam}$ tennessee]]’

CITIZEN-OR-RESIDENT(‘gore’, ‘gore’s home state of tennessee’)
CITIZEN-OR-RESIDENT(‘gore’, ‘tennessee’).

Rule 4 maps to any coreferent and embedding entity mention, for example:

‘$[^{gpl}_{nam}$ $[^{gpl}_{nom}$ the $[^{gpl}_{nom}$ West African] nation] of Senegal]’,

PART-WHOLE(‘the West African nation’, ‘West African’),
PART-WHOLE(‘the West African nation of Senegal’, ‘West African’).

https://doi.org/10.1017/S1351324911000106 Published online by Cambridge University Press

Table 4. *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 percentage of firings for ACE 2004 and ACE 2005*

#	Description of mapping rule			ACE 2004	ACE 2005
1	Prenominal	↦	Embedding coreferent	191 (26.6%)	104 (20.6%)
2	Nominal	↦	Embedded left adjacent prenominal	30 (4.2%)	34 (6.7%)
3	Nominal	↦	Embedded coreferent	41 (5.7%)	73 (14.5%)
4	Nominal	↦	Embedded coreferent	11 (1.5%)	3 (0.6%)
5	Nominal	↦	Left adjacent coreferent	178 (24.8%)	69 (13.7%)
6	Nominal	↦	Right adjacent coreferent	133 (18.5%)	106 (21.0%)
7	Nominal	↦	Left adjacent coreferent, skip copula	35 (4.9%)	31 (6.2%)
8	Nominal	↦	Right adjacent coreferent, skip copula	3 (0.4%)	1 (0.2%)
9	Nominal	↦	Left adjacent coreferent, skip Verb+TO_BE	1 (0.1%)	1 (0.2%)
10	Nominal	↦	Right adjacent coreferent, Skip verb+TO_BE	0 (0.0%)	0 (0.0%)
11	Nominal	↦	Left adjacent coreferent, skip copula and coreferring entities	3 (0.4%)	1 (0.2%)
12	Nominal	↦	Right adjacent coreferent, skip copula and coreferring entities	1 (0.1%)	0 (0.0%)
13	Nominal	↦	Left coreferent	89 (12.4%)	73 (14.5%)
14	Nominal	↦	Right coreferent	3 (0.4%)	8 (1.6%)

Note that this relation can be further simplified to PART-WHOLE('Senegal', 'African') by taking entity heads only.

Rules 5 and 6 map to immediately adjacent coreferent entity mentions. These rules, as described above, are illustrated in Figure 14(b) and (c). Rules 7 and 8 map to coreferent entity mentions found respectively to the left or the right, and 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, for example:

'$[^{per}_{nom}$ 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 again map to coreferent entity mentions found respectively to the left or the right. However, they allow two intervening verb phrases and any number of adverbs as long as the second verb phrase consists of an infinitival copula (i.e. 'to be'), for example:

'$[^{per}_{nam}$ 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.').

https://doi.org/10.1017/S1351324911000106 Published online by Cambridge University Press

Rules 11 and 12 map once again to coreferent entity mentions found respectively to the left or 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, for example:

'[$_{nam}^{per}$ Card] is [$_{nom}^{per}$ a [$_{nam}^{gpl}$ Washington] insider] and [$_{nom}^{per}$ a lobbyist for [$_{nam}^{org}$ 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 lobbyist for 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 the right, for example:

'[$_{nom}^{org}$ [$_{nam}^{per}$ martha stewart]'s company], officially known as [$_{nam}^{org}$ m. s. living omnimedia]'

EMPLOY-EXECUTIVE('martha stewart', 'martha stewart's company')
EMPLOY-EXECUTIVE('martha stewart', 'm. s. living omnimedia').

Rules 13 and 14 fired a total of ninety-two times for the ACE 2004 data and eighty-two times for the ACE 2005 data (Table 4). In order to assess the accuracy of Rules 13 and 14, a random sample of twenty firings was inspected for ACE 2004. Six (30%) were found to be noisy (details below), which correspond to 27.6% of the full ninety-two firings of Rules 13 and 14. Since Rules 1–12 do not introduce noise, the overall error rate for the ACE 2004 rule firings listed in Table 4 is 3.8% (27.6/719), and the overall error rate for the full ACE 2004 data set used in the experiments here (see Table 7) is less than 2% (27.6/1,400). We consider here a small amount of noise to be a reasonable tradeoff for a larger data set.

Among the sample, four firings create relation mentions that are questionable or could arguably have been mapped to a more suitable target entity mention, e.g. Rule 13 triggers the following mapping:

'[$_{nam}^{per}$ Ehud Barak] won the endorsement of [$_{nom}^{org}$ [$_{pro}^{per}$ his] Labor party] as [$_{nom}^{per}$ [$_{pro}^{org}$ 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'. However, mapping away from pronominal mention here is inconsistent with the overall goal of a corpus of relations over named and pronominal entity mentions.

Another noisy mapping is encountered when a nominal entity mention is mapped out of an embedded entity mention, but the other entity mention is a possessive pronoun that is not mapped, for example:

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

BUSINESS('their', 'their candidate')
BUSINESS('their', 'Gore').

https://doi.org/10.1017/S1351324911000106 Published online by Cambridge University Press

Table 5. *List of rules for reannotation of in BioInfer. Columns contain the rule identifier (#), the rule description and the affected relation types (description/relation type) and the number and percentage of relation mentions of the given type in the source (source) and the mapped (mapped) data*

#	Description/relation type			Source		Mapped	
(1)	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%)
(2)	*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%)

This questionable mapping occurred twice in the sample. In the current work, the mapping is allowed to fire and the resulting relation mentions are kept in the final data.

6.2 Reannotating BioInfer

The reannotation of BioInfer is motivated by the fact that the BioInfer annotation sometimes marks part-whole and part-part relation mentions differently depending on their syntactic context. It is also motivated by the fact that the BioInfer annotation sometimes marks relations with more than two arguments. Table 5 contains details of the mapping rules. 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. Part-Whole ↦ Part-Part). 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 are mapped and the corresponding percentage (of the total number of mapped relation mentions). The fourth column (Mapped) 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 in Table 5 addresses relation mentions that are marked differently depending on their syntactic context, by mapping part-whole relation mentions to part-part. 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.'

https://doi.org/10.1017/S1351324911000106 Published online by Cambridge University Press

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 (Pyysalo *et al.* 2007), BIND is defined as a non-covalent binding (i.e. formation of a complex, association) between the arguments and CONTAIN is defined as a component being part of a complex. For the annotation to be consistent across the two sentences, the second sentence should also have a relation mention 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'). Co-*X* relations do not have a clear semantics; however, they do exist and are expressed in the text. Therefore, they are used here for the relation identification experiments (Section 8.1) but ignored for the relation characterisation experiments (Section 8.2).

Rule 2 in Table 5 addresses relations over more than two entity mentions in BioInfer, by mapping n-ary relation mentions to binary relation mentions over all possible entity mention pairs. Consider the following example:

'Immediately after synthesis, [^{protein} E-cadherin], [^{protein} beta-catenin], and [^{protein} plakoglobin] cosedimented as complexes.'

MUTUALCOMPLEX('E-cadherin', 'beta-catenin', 'plakoglobin'),

MUTUALCOMPLEX('E-cadherin', 'beta-catenin'),
MUTUALCOMPLEX('E-cadherin', 'plakoglobin'),
MUTUALCOMPLEX('beta-catenin', 'plakoglobin').

Here, the top relation mention with three arguments is replaced by the three distinct binary relation mentions that follow it. Note that it has been argued that biomedical relations need to be n-ary (Rzhetsky *et al.* 2004; Wattarujeekrit, Shah and Collier 2004; McDonald *et al.* 2005; Cohen and Hunter 2006). However, while the mapped binary relation mentions here do not necessarily capture the simultaneous interaction expressed in the original annotation, they are appropriate for many applications (e.g. large-scale automatic search and knowledge discovery tasks like GRE) and are compatible with the domain-general notion of relation adopted here. Furthermore, of 2,424 relations that result from the refactoring described in Section 3.2, only thirty-four are n-ary relations (Table 5). In addition, the original relation identifiers are retained, allowing n-ary relations to be reconstructed.

7 Example usage: generic relation extraction (GRE)

The goal of GRE is to identify mentions of relations in text using techniques that achieve comparable accuracy when transferred across domains without the modification of model parameters. Applications of GRE include (1) interactive knowledge discovery (e.g. where new relations are fed to a human analyst through a visualisation tool that indicates strengths and types of associations); (2) initialisation of the RE bootstrapping (e.g. where relation clusters are used to intialise active learning on

https://doi.org/10.1017/S1351324911000106 Published online by Cambridge University Press

previously unseen data) and (3) noisy knowledge representation (e.g. for applications such as paraphrase acquisition and automatic summarisation (Hasegawa et al., 2005; Hachey, 2009a).

The leftover subsections describe the final data preparation for the GRE experiments here. The goal is to derive data sets that are comparable in terms of the total number of relation mentions, the number of subsets and the number of relation types per subset. First, entity mentions are filtered where possible, keeping only those that refer to specific objects and are names or pronouns. Second, relation mentions are filtered, keeping only those that describe real-world relationships. Third, the entity and relation-type schemas are converted, merging some classes to avoid sparseness in the final data sets.

Furthermore, relation mentions are required to be between two entity mentions that are in the same sentence[16] and are distinct siblings. The requirement that the entity mentions be *distinct* removes reflexives, which are relation mentions where either both entity mentions are identical or the type 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 as being coreferent, mapping rule 4 (described in Section 6.1) fires and the relation mention ends up being SUBSIDIARY('afghanistan','afghanistan'). Moreover, 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 in the entity mention constituent tree) are not considered. Consider the following relation mentions from the text snippet 'E-cadherin/plakoglobin complexes':

CHANGE/PHYSICAL('E-cadherin', 'plakoglobin'),
OBJECT-COMPONENT('E-cadherin/plakoglobin complexes', 'E-cadherin'),
OBJECT-COMPONENT('E-cadherin/plakoglobin complexes', 'plakoglobin').

Here, the CHANGE/PHYSICAL('E-cadherin', 'plakoglobin') relation mention is kept, but the other two relation mentions are ignored. The *sibling* requirement also means that other long-distance relationships within the entity mention constituent tree (e.g. cousins) are not considered.

Finally, seven entity pair subsets are chosen for each data set based on two sparseness criteria. First, relation types are considered to be outliers and filtered if they have less than three total mentions. Second, entity pair domains are only used for generic relation characterisation (GRC) if they have thirty or more total

[16] 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.

https://doi.org/10.1017/S1351324911000106 Published online by Cambridge University Press

mentions and two or more distinct relation types. Relation distributions for the resulting data sets are given in the respective subsections below.

7.1 ACE data for GRE

Filtering entity mentions. 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.) This serves to remove all nominal mentions, which are not reliably recognised by most named entity recognition systems. Prenominal mentions are kept because they are often names (e.g. '<u>Labour</u>' in '[$^{per}_{nom}$ [$^{org}_{pre}$ Labour] nominee]'), though not always (e.g. 'British <u>prime minister</u>' 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. Entity mentions are also filtered based on mention class. The filtering here removes all entity mentions that are underspecified (e.g. 'many people'), generic (e.g. 'extremist groups') or negatively quantified (e.g. 'no one'). This leaves only entity mentions that are specific referential, i.e. mentions that refer to a particular, unique object or set of objects in the real world.

Filtering relation mentions. Next, relation mentions are removed where one of the entity mentions is no longer part of the annotation because of 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 2004b), *'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 13,358 entity mentions and 1,511 relation mentions (down from 22,736 and 4,374, respectively, in the original source). And the ACE 2005 data has 10,345 entity mentions and 975 relation mentions (down from 18,086 and 3,658, respectively).

Converting to the final schema. Finally, 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 6 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, respectively. In the Source and Target columns, entity and relation-type labels prefixed with 'T:' are types and labels prefixed with 'S:' are subtypes. Rows 1 through 2 of Table 6(a), for example specify that all entity mentions with type GPE (geo-political) or LOC (location) are changed to have a single common type GPL. And, Rows 1 through 4 of

https://doi.org/10.1017/S1351324911000106 Published online by Cambridge University Press

Table 6. *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 subtypes, respectively*

#	Source	Target	ACE 2004		ACE 2005	
		(a) Entity-type changes				
(1)	T:GPE (Geo-Political)	T:GPL	3,262	(87.5%)	3,330	(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	(0.7%)
		(b) Relation-type changes				
(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		NA	NA	49	(8.5%)
(5)	S:Cit-Res,	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 & S:Other,		15	(1.9%)	NA	NA
(9)	S:Cit-Res-Rel-Eth		NA	NA	42	(7.3%)
(10)	T:ART	T:AGT-ART & S:Use-Own-Inv-Mnf	14	(1.8%)	35	(6.1%)
(11)	S:Subsidiary	T:PRT-WHL & S:Subsidiary	80	(10.2%)	81	(14.0%)
(12)	S:Part-Whole,	T:PRT-WHL	187	(23.9%)	NA	NA
(13)	T:PART-WHOLE		NA	NA	137	(23.7%)

Table 6(b) specify that all relation mentions with subtype LOCATED, NEAR, BASED-IN or ORG-LOCATION (located, based, headquartered, operates etc.) are changed to have type GEN-AFF (affiliation) and subtype LOCATED.

The GRE data set. Tables 7 and 8 contain the generic relation identification (GRI)- and GRC-type distributions for the final ACE 2004 and ACE 2005 data sets as used for the experiments here. The first column lists the gold standard type. For GRI, this is a binary distinction between a pair of entity mentions being in a relation or not being in a relation. Possible entity pairs are all those occurring within the same sentence. For GRC, the first column lists the relation type (with supertypes typeset in small capital letters). The next seven columns list the entity pair subdomains. 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). Note that for ACE, the the number of GRI Y instances is less than the total number of GRC instances because relation mentions are removed where one of the entity mentions are prenominal (e.g. 'Scottish' in 'Scottish National Health Service'). This is to make the GRI task

https://doi.org/10.1017/S1351324911000106 Published online by Cambridge University Press

Table 7. *Relation distributions for GRE news development data (ACE 2004). The first column specifies the relation type and the following columns specify the entity pair subdomains*

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	1,041	749	1,805	756	1,480	2,408
Total	91	1,200	841	2,071	798	1,788	2,464
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-MANUFACTURER	6	–	–	–	–	–	–
Total	32	189	164	401	51	509	54

consistent with the named entity recognisers used for a related extrinsic evaluation (Hachey 2009a), which do not mark prenominal entity mentions. These instances are not filtered for the GRC data in order to maximise the number of data points for GRC evaluation.

7.2 BioInfer data for GRE

Filtering entity mentions. First, we only consider relations between PHYSICAL entities, defined as *a physical, biochemical object[s]* (Ginter *et al.* 2007). While BioInfer annotates relations between PHYSICAL entities, properties and processes, the experiments here focus on the subtask of relations between PHYSICAL entities. Therefore, BioInfer entity mentions with PROPERTY and PROCESS supertypes are ignored. We also ignore TEXTBINDING entity mentions, which are used to mark spans of text that express a relation (example in Figure 5). In addition, we filter

https://doi.org/10.1017/S1351324911000106 Published online by Cambridge University Press

Table 8. *Relation distributions for GRE news test data (ACE 2005). The first column specifies the relation type and the following columns specify the entity pair subdomains*

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	1,216	658	1,405	914	1,149
Total	117	59	1,303	692	1,606	1,033	1,210
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

redundant annotations of the same entity mention. This can happen, for example, when a plural pronoun 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 '4b'. Here, mentions that do not take part in a relation are removed until only one is left.

Filtering relation mentions. Next, relation mentions are removed where one of the entity mentions is no longer part of the annotation because of the entity filtering rules. Relation mentions are also filtered based on type. In particular, relation mentions with type REL-ENT are removed. These are BioInfer relations where an unnamed entity mention refers to a named entity mention, for example:

https://doi.org/10.1017/S1351324911000106 Published online by Cambridge University Press

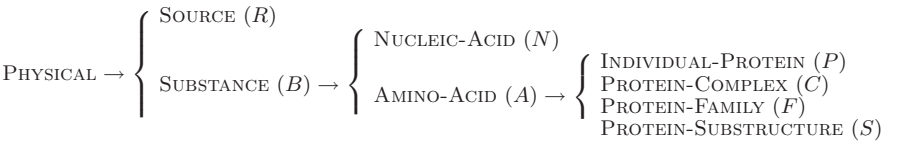


Fig. 15. Simplified entity-type schema for BioInfer.

'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 fifty Rel-Ent relation mentions. After filtering, the BioInfer data has 5,800 entity mentions and 2,116 relation mentions (down from 7,818 and 3,020, respectively, in the original source).

Converting to the final schema. Finally, entity and relation types are changed to the final schema. For BioInfer, this is a matter of choosing a level in the full relation-type schema from the source data that gives several entity pair subdomains with a sufficient number of relation types and instances for evaluation of the GRC task. Figure 15 contains a simplified version of the 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 subdomain 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 subdomain would be R-B. The relation type for the GRC task is simply the second-level type from the full-relation schema (Pyysalo *et al.* 2007), i.e. one of Causal, Part-Of, Observation or Is-A.

The GRE data set. Table 9 contains the GRI- and GRC-type distributions for the final BioInfer data set as used for the experiments here. 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 supertypes typeset in small capital letters). The next seven columns list the entity pair subdomains. Note that the number of instances in the GRC data subsets is less than the corresponding number of GRI Y instances because the number of relation mentions that have vague or undetermined types are ignored for the relation characterisation experiments (but not the relation identification experiments). These consist of Corelate (a general, unspecified relation between the arguments), Humanmade (a relationship that is forced or caused by human intervention), Relate (a general, unspecified, non-directional relationship used when no details of the relationship are known) and Co-* (the relations created by Rule 1 for mapping BioInfer entity mentions from Section 6.2).

https://doi.org/10.1017/S1351324911000106 Published online by Cambridge University Press

Table 9. *Relation distributions for GRE biomedical test data (BioInfer). The first column specifies the relation type and the following columns specify the entity pair subdomains*

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	2,450	183	521	229	362	325
Total	225	3,392	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

8 GRE experiments

This section contains overview experiments' comparing approaches to GRE across domains. Models are tuned on the news development data (ACE 2004) and tested both on news test data (ACE 2005) and biomedical test data (BioInfer). Results validate the GRE claim of modification-free adaptation. For system details, full experiments and error analysis, refer to Hachey (2009b). For an extrinsic evaluation exploring the utilitiy of end-to-end GRE for automatic summarisation, refer to Hasegawa *et al.* (2005) or Hachey (2009a).

8.1 Experiment I: portability of generic relation identification (GRI)

The first step in GRE is GRI, where the goal is to identify relation-forming entity mention pairs using methods that port across domains without modification of model parameters. The input to the GRI task consists of sentences from source documents with entity mention markup. For the purpose of the intrinsic evaluation here, gold standard entity mention annotation is used, serving to isolate the errors that are due to the GRI module. All pairs of entity mentions that occur in the same sentence are considered to be the candidate relation mentions. Only considering intrasentential 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 one instance of a gold standard relation mention where the entity mentions are in different sentences. 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.

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

$$P = \frac{Num\ correct}{Total\ system\ pairs} \qquad R = \frac{Num\ correct}{Total\ gold\ pairs},$$

https://doi.org/10.1017/S1351324911000106 Published online by Cambridge University Press

and, f-score (F) is calculated in the standard way: $F = 2PR/(P + R)$. Paired Wilcoxon signed ranks tests across entity pair subdomains are used to check for significant differences between systems. The paired Wilcoxon signed ranks test is a non-parametric analogue of the paired t test. The null hypothesis is that the two populations from which the scores are sampled are identical. Following convention, the null hypothesis is rejected for values of p less than or equal 0.05. Subdomains are formed by taking just those relations between two entities of given types (details given in Tables 7, 8 and 9).

8.1.1 Systems

Atomic events (*Event*). The first model of entity mention co-occurrence is based on an approach from the literature for identifying atomic events (Filatova and Hatzivassiloglou 2003). 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.

Intervening token windows (*Toks*). The next model is based on intervening token windows. It accepts all pairs of entity mentions that (1) occur in the same sentence, and (2) have t or fewer intervening tokens. Most previous GRI works have used some variant of this model. Hasegawa *et al.* (2004), for example, use this approach but do not motivate their threshold of $t = 5$. Based on tuning experiments on the news development data (ACE 2004), the threshold here is set to $t = 2$.

Dependency path windows (*Deps*). The experiments here also consider a novel approach to modelling entity mention co-occurrence that is based on syntactic governor-dependency relations. This accepts all pairs of entity mentions that (1) occur in the same sentence, and (2) have d or fewer intervening token nodes on the shortest dependency path connecting the two entity mentions. Note that a further collapsing of dependency paths is performed here that passes governor-dependency relations along chains of conjoined tokens in the intervening context. So, for example, the path between 'Murray' and 'Awadi' in Figure 10 (Section 5.1 above) has one intervening token node ('recruited') instead of having two ('recruited' and 'Berry'). Based on tuning experiments on the news development data (ACE 2004), the threshold here is set to $d = 0$.

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

8.1.2 Results

Table 10 contains P, R and F results. The best score for each measure is in bold and scores that are statistically distinguishable from the best ($p \leq 0.05$) are underlined.

https://doi.org/10.1017/S1351324911000106 Published online by Cambridge University Press

Table 10. *Comparison of P, R and F on news and biomedical test sets. The best score in each column is in bold and those that are statistically distinguishable from the best are underlined*

	ACE 2005 (News test set)			BioInfer (Biomedical test set)		
	P	*R*	*F*	*P*	*R*	*F*
Baseline	0.110	1.000	0.195	0.268	1.000	0.415
Event	0.050	0.392	0.083	0.186	0.418	0.247
Toks	0.291	0.510	0.342	**0.527**	0.388	0.422
Deps	**0.456**	0.392	**0.360**	0.450	0.302	0.349
Comb	0.277	**0.538**	0.332	0.500	**0.454**	**0.453**
Human	0.906	0.675	0.773	NA	NA	NA

The first row corresponds to a baseline system that accepts all pairs of entity mentions occurring in the same sentence. The final row corresponds to the upper bound calculated in terms of mean human agreement with respect to the adjudicated gold standard. Results suggest that the 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).

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

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

https://doi.org/10.1017/S1351324911000106 Published online by Cambridge University Press

(1.000), this is due to higher precision (0.268 as opposed to 0.110), which is due to the higher proportion of true relation-forming pairs (27% for BioInfer compared to approximately 10% for the ACE data sets). This is artificially high, however, since the BioInfer creators selectively sampled sentences that include mentions of proteins that are known to interact. Comb precision is significantly better than the Baseline precision on both domains.

8.2 Experiment II: portability of generic relation characterisation (GRC)

The second step in GRE is GRC, where the goal is to automatically annotate each relation mention with a label that describes the relation type using methods that port across domains without modification of model parameters. The input to the GRC task is the output from the GRI task 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, serving to isolate the errors that are due to the GRC module. The primary modelling task of GRC is to induce 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 task is an instance for clustering. In other words, the clustering instance level is entity pair tokens instead of entity pair types.

Accuracy is measured in terms of two different approaches to calculating precision (P), recall (R) and f-score (F) of clustering output with respect to a gold standard. $F_{1:1}$ is based on a one-to-one mapping between clusters and gold standard classes. If there are more classes than clusters, then some classes are left unaligned (likewise if there are more clusters). Calculating the one-to-one mapping can be expensive; 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. F_{pw} is based on the number of pairs of data points that are in the same cluster and in the same class (tp), the number of pairs in the same cluster but in different classes (fp) and the number of pairs in different clusters but in the same class (fn). As above, the paired Wilcoxon signed ranks tests across entity pair subdomains are used to check for significant differences between systems.

8.2.1 Systems

Unreduced. In the second approach, no dimensionality reduction is performed. Here feature vectors are extracted for each relation mention and weighted using *tf*idf*, which is calculated as follows:

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

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

https://doi.org/10.1017/S1351324911000106 Published online by Cambridge University Press

which feature i occurs. Cosine is used to measure the similarity between feature vectors in the unreduced feature space.

SVD-reduced. In the third approach, dimensionality reduction is performed using singular value decomposition (SVD), a linear algebraic least squares method (Eckart and Young 1936). In case where $X_{r\times f}$ is a *tf*idf*-weighted relation-by-feature (R×F) matrix, 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. Based on tuning experiments on the ACE 2004 data, n is set to 5 for the experiments here.

LDA-reduced. While SVD has proved successful, its representation of words and documents (or relations) as points in a Euclidean space is not easy to interpret. In the fourth approach, dimensionality reduction is performed using Latent Dirichlet allocation (LDA) (Blei *et al.* 2003), a generative probabilistic version of latent semantic analysis (Berry et al. 1995; Landauer, Foltz and Laham 1998). 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. In case where T is the number of topics, the probability of the ith feature is written as follows:

$$P(f_i) = \sum_{j=1}^{T} P(f_i|z_i = j)P(z_i = j), \tag{2}$$

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 the features that are important to a topic and $P(\mathbf{z})$ is the prevalence of those topics for a given relation mention (Griffiths and Steyvers 2004). The resulting similarity model contains four free parameters: the number of topics T, two hyperparameters (β and α, which determine the nature of the Dirichlet priors on $P(f_i|z_i = j)$ and $P(z_i = j)$, respectively) and the constant C for divergence-to-similarity conversion. Based on tuning experiments on the ACE 2004 data, these are set to $T = 0.97|TotalFeatures|$, $\beta = 0.0001$, $\alpha = 50/T$ and $C = 8$ for the experiments here.

8.2.2 Results

Table 11 contains P, R and F results. The best score for each measure is in bold and scores that are statistically distinguishable from the best ($p \leq 0.05$) are underlined. The first row corresponds to a baseline system that randomly partitions the data into n clusters. The final row corresponds to the upper bound calculated in terms of

https://doi.org/10.1017/S1351324911000106 Published online by Cambridge University Press

Table 11. *Comparison of precision, recall and f-score results on all data sets*

	$P_{1:1}$	$R_{1:1}$	$F_{1:1}$	P_{pw}	R_{pw}	F_{pw}
Baseline	0.583	0.357	0.437	0.521	0.295	0.372
Unreduced	0.720	0.511	0.591	**0.616**	0.414	0.486
SVD-reduced	**0.726**	0.540	0.609	**0.616**	0.414	0.486
LDA-reduced	0.692	**0.685**	**0.683**	0.551	**0.923**	**0.676**
Human	0.969	0.923	0.966	0.946	0.937	0.941

(a) ACE 2004 (News development set)

	$P_{1:1}$	$R_{1:1}$	$F_{1:1}$	P_{pw}	R_{pw}	F_{pw}
Baseline	0.414	0.429	0.485	0.509	0.366	0.415
Unreduced	**0.674**	0.566	**0.607**	**0.552**	0.511	0.513
SVD-reduced	0.663	0.555	0.599	0.543	0.523	0.518
LDA-reduced	0.564	**0.634**	0.591	0.523	**0.875**	**0.646**
Human	0.969	0.923	0.966	0.946	0.937	0.941

(b) ACE 2005 (News test set)

	$P_{1:1}$	$R_{1:1}$	$F_{1:1}$	P_{pw}	R_{pw}	F_{pw}
Baseline	0.655	0.444	0.525	0.597	0.374	0.455
Unreduced	0.729	0.522	0.600	**0.644**	0.457	0.526
SVD-reduced	**0.765**	0.596	0.663	0.639	0.586	0.587
LDA-reduced	0.720	**0.705**	**0.708**	0.606	**0.779**	**0.672**
Human	0.969	0.923	0.966	0.946	0.937	0.941

(c) BioInfer (Biomedical test set)

mean human agreement with respect to the adjudicated gold standard. All clustering approaches here use a feature set consisting of (1) words occurring between the two entities, (2) words occurring within the entity phrases and (3) words and grammatical relations occurring on the path connecting the two entities in the dependency parse (Hachey 2009b). 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 11(a) contains results for the news domain development set (ACE 2004); Table 11(b) contains results for the news domain test set (ACE 2005); and Table 11(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 data sets and evaluation measures. Moreover, 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. 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. The effect of the hyperparameters can be observed

https://doi.org/10.1017/S1351324911000106 Published online by Cambridge University Press

in the relatively high recall for the LDA-reduced model. Here the small values of α (means across subdomains of 0.63, 0.67 and 0.65, respectively, for the ACE 2004, ACE 2005 and BioInfer) 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).

9 Conclusion

This paper discussed data sets for multi-type relation extraction across domains. We defined relations as associations between named or pronominal entities. Furthermore, we specified that relations are between exactly two entities and part-whole and part-part relations should be consistently marked. The result is a common notion of relation that serves as a middle ground between different RE corpora. In addition, this notion of a relation is compatible with the semantic web and the linked data movements and with large-scale, automatic search and knowledge discovery tasks (e.g. GRE).

Two standard and publicly available RE corpora were adapted to comply with this definition via a three-stage process (refactoring, pre-processing and reannotation). The ACE 2004 and 2005 corpora were used to derive news data and the BioInfer corpus was used to derive biomedical data. These corpora were chosen because they have multiple relation types that are not determined by the types of the participating entities.

Finally, we reported experiments for relation identification and characterisation to illustrate the application and utility of these corpora. The experiments used three comparable data sets: (1) the ACE 2004 data for development in the news domain; (2) the ACE 2005 data for testing in the news domain; and (3) the BioInfer data for testing in the biomedical domain. This allowed evaluation across distinct epochs within the news domain, validating the GRE claim of modification-free domain adaptation.

References

Agichtein, E., and Gravano, L. 2000. Snowball: extracting relations from large plain-text collections. In *Proceedings of the 5th ACM Conference on Digital Libraries*, pp. 85–94. New York, NY: ACM.

Aone, C., Halverson, L., Hampton, T., and Ramos-Santacruz, M. 1998. SRA: description of the IE2 system used for MUC-7. In *Proceedings of the 7th Message Understanding Conference* (MUC-7), Columbia, MD. Gaithersburg: NIST.

Auer, S., Dietzold, S., Lehmann, J., Hellmann, S., and Aumueller, D. 2009. Triplify: light-weight linked data publication from relational databases. In *Proceedings of the 18th International World Wide Web Conference*, Madrid, Spain, pp. 621–30. New York, NY: ACM.

Berry, M. W., Dumais, S. T., and O'Brien, G. W. 1995. Using linear algebra for intelligent information retrieval. *SIAM Review* **37**(4): 573–95.

Bizer, C., Heath, T., and Berners-Lee, T. 2009. Linked data – the story so far. *International Journal on Semantic Web and Information Systems* **5**(3): 1–22.

https://doi.org/10.1017/S1351324911000106 Published online by Cambridge University Press

Blei, D., Ng, A. Y., and Jordan, M. I. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* **3**: 993–1022.

Brin, S. 1999. Extracting patterns and relations from the world wide web. In: P. Atzeni, A. Mendelzon, and G. Mecca (eds.), *The World Wide Web and Databases: Selected Papers from WebDB '98*, pp. 172–83. Lecture Notes in Computer Science. Berlin: Springer.

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): 139–55.

Byrne, K. 2009. *Populating the Semantic Web – Combining Text and Relational Databases as RDF Graphs*. PhD thesis, University of Edinburgh.

Chinchor, N. 1998. Overview of MUC-7. In *Proceedings of the 7th Message Understanding Conference*. Gaithersburg, MD: NIST.

Cohen, K. B., Fox, L., Ogren, P. V., and Hunter, L. 2005. Corpus design for biomedical natural language processing. In *Proceedings of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics*, pp. 38–45. Morristown, TN: ACL.

Cohen, K. B., and Hunter, L. 2006. A critical review of PASBio's argument structures for biomedical verbs. *BMC Bioinformatics* **7**(Suppl 3): S6.

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*, pp. 260–70. New York, NY: ACM.

Curran, J. R., and Clark, S. 2003. Investigating GIS and smoothing for maximum entropy taggers. In *Proceedings of the 11th Meeting of the European Chapter of the Association for Computational Linguistics*, pp. 91–8. Morristown, TN: ACL.

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*, pp. 837–40. Paris: ELDA.

Eckart, C., and Young, G. 1936. The approximation of one matrix by another of lower rank. *Psychometrika* **1**(3): 211–218.

Filatova, E., and Hatzivassiloglou, V. 2003. Marking atomic events in sets of related texts. In: N. Nicolov, K. Bontcheva, G. Angelova, and R Mitkov (eds.), *Recent Advances in Natural Language Processing III*, pp. 247–56. Amsterdam, Netherlands: John Benjamins.

Ginter, F., Pyysalo, S., Björne, J., Heimonen, J., and Salakoski, T. 2007. BioInfer relationship annotation manual. Technical Report 806, Turku Centre for Computer Science.

Griffiths, T. L., and Steyvers, M. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences* **101**(Suppl 1): 5228–5235.

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*, pp. 19–26. Morristown: ACL.

Hachey, B. 2009a. Multi-document summarisation using generic relation extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Singapore, pp. 420–9. Morristown, TN: ACL.

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

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*, pp. 415–22. Morristown, TN: ACL.

Hasegawa, T., Sekine, S., and Grishman, R. 2005. Unsupervised paraphrase acquisition via relation discovery. Technical Report 05-012, Proteus Project, Computer Science Department, New York University.

Heimonen, J., Pyysalo, S., Ginter, F., and Salakoski, T. 2008. Complex-to-pairwise mapping of biological relationships using a semantic network representation. In *Proceedings of the*

https://doi.org/10.1017/S1351324911000106 Published online by Cambridge University Press

3rd International Symposium on Semantic Mining in Biomedicine, pp. 45–52. Turku: Turku Centre for Computer Science Turku, Finland.

Johnson, H. L., Jr., William A. Baumgartner, Krallinger, M., Cohen, K. B., and Hunter, L. 2007. Corpus refactoring: a feasibility study. *Journal of Biomedical Discovery and Collaboration* **2**: 4.

Landauer, T. K., Foltz, P. W., and Laham, D. 1998. An introduction to latent semantic analysis. *Discourse Processes* **25**(2): 259–284.

Linguistic Data Consortium (LDC) 2004a. *Annotation Guidelines for Entity Detection and Tracking (EDT)*. Philadelphia, PA: LDC. `http://www.ldc.upenn.edu/Projects/ACE/docs/EnglishEDTV4-2-6.PDF` Accessed 22 July 2008.

Linguistic Data Consortium (LDC) 2004b. *Annotation Guidelines for Relation Detection and Characterization (RDC)*. Philadelphia, PA: LDC. `http://www.ldc.upenn.edu/Projects/ACE/docs/EnglishRDCV4-3-2.PDF`. Accessed 22 July 2008.

Linguistic Data Consortium (LDC) 2005a. *ACE (Automatic Content Extraction) English Annotation Guidelines for Entities*. Philadelphia, PA: LDC. `http://www.ldc.upenn.edu/Projects/ACE/docs/English-Entities-Guidelines_v5.6.1.pdf`. Accessed 22 July 2008.

Linguistic Data Consortium (LDC) 2005b. *ACE (Automatic Content Extraction) English Annotation Guidelines for Relations*. Philadelphia, PA: LDC. `http://www.ldc.upenn.edu/Projects/ACE/docs/English-Relations-Guidelines_v5.8.3.pdf`. Accessed 22 July 2008.

Lin, D. 1998. Dependency-based evaluation of MINIPAR. In *Proceedings of the LREC Workshop Evaluation of Parsing Systems*, pp. 317–30. Paris: ELDA.

Lin, D., and Pantel, P. 2001. Discovery of inference rules for question answering. *Natural Language Engineering* **7**(4): 343–360.

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. 1993. Building a large annotated corpus of English: the Penn treebank. *Computational Linguistics* **19**(2): 313–30. ISSN 0891-2017.

McDonald, R., Pereira, F., Kulick, S., Winters, S., Jin, Y., and White, P. 2005. Simple algorithms for complex relation extraction with applications to biomedical IE. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pp. 491–8. Morristown, TN: ACL.

Minnen, G., Carroll, J., and Pearce, D. 2000. Robust, applied morphological generation. In *Proceedings of the 1st International Natural Language Generation Conference*, pp. 201–8. Morristown, TN: ACL.

Mitchell, A., Strassel, S., Huang, S., and Zakhary, R. 2005. *ACE 2004 Multilingual Training Corpus*. Philadelphia, PA: Linguistic Data Consortium.

Pustejovsky, J., Saurí, R., Castaño, J., Radev, D., Gaizauskas, R., Setzer, A., Sundheim, B., and Katz, G. 2004. Representing temporal and event knowledge for QA systems. In: M. T. Maybury (ed.), *New Directions in Question Answering*, pp. 99–112. Menlo Park, CA: AAAI Press.

Pyysalo, S., Airola, A., Heimonen, J., Björne, J., Ginter, F., and Salakoski, T. 2008. Comparative analysis of five protein–protein interaction corpora. *BMC Bioinformatics* **9**(Suppl 3): S6.

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.

Rzhetsky, A., Iossifov, I., Koike, T., Krauthammer, M., Kra, P., Morris, M., Yu, H., Dubou, P. A., Weng, W., Wilbur, W. J., Hatzivassiloglou, V., and Friedman, C. 2004. Geneways: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. *Journal of Biomedical Informatics* **37**(1): 43–53.

Sekine, S. 2006. On-demand information extraction. In *Proceedings of the COLING/ACL Main Conference Poster Sessions*, pp. 731–8. Morristown, TN: ACL.

https://doi.org/10.1017/S1351324911000106 Published online by Cambridge University Press

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*, pp. 73–80. New York, NY: ACM.

Swanson, D. R. (1986) Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine* **30**(1): 7–18.

Turmo, J., Ageno, A., and Català, N. 2006. Adaptive information extraction. *ACM Computing Surveys*, **38**(2): 4.

Walker, C., Strassel, S., Medero, J., and Maeda, K. 2006. *ACE 2005 Multilingual Training Corpus*. Philadelphia, PA: Linguistic Data Consortium.

Wattarujeekrit, T., Shah, P., and Collier, N. 2004. PASBio: predicate-argument structures for event extraction in molecular biology. *BMC Bioinformatics* **5**: 155.

https://doi.org/10.1017/S1351324911000106 Published online by Cambridge University Press