Inter-sentential Relations in Information Extraction Corpora

Kumutha Swampillai, Mark Stevenson

Department of Computer Science,
University of Sheffield,
Regent Court, 211 Portobello,
Sehffield, S1 4DP
United Kingdom
{k.swampillai, m.stevenson}@dcs.shef.ac.uk

Abstract

In natural language relationships between entities can asserted within a single sentence or over many sentences in a document. Many information extraction systems are constrained to extracting binary relations that are asserted within a single sentence (single-sentence relations) and this limits the proportion of relations they can extract since those expressed across multiple sentences (inter-sentential relations) are not considered. The analysis in this paper focuses on finding the distribution of inter-sentential and single-sentence relations in two corpora used for the evaluation of information extraction systems: the MUC6 corpus and the ACE corpus from 2003. In order to carry out this analysis we had to manually mark up all the management succession relations described in the MUC6 corpus. It was found that inter-sentential relations constitute 28.5% and 9.4% of the total number of relations in MUC6 and ACE03 respectively. This places upper bounds on the recall of information extraction systems that do not consider relations that are asserted across multiple sentences (71.5% and 90.6% respectively).

1. Introduction

Relation extraction is a subtask of Information Extraction (IE) that aims to identify instances of pre-defined relation within text. For example, the sentence *"John Scheurer was appointed as the new CEO of Allied Capital last week."* asserts a relation between *"John Scheurer"* and *"Allied Capital"*.

Many relation extraction systems constrain the search for relations to ones that are asserted within a single sentence, for example (Zelenko et al., 2003; Culotta and Sorensen, 2004; Zhao and Grishman, 2005; Zhou et al., 2007). We refer to these as single-sentence relations. The practical advantage of this approach is that it limits both the computational complexity of the task and data sparsity issues that arise with machine learning approaches to relation extraction. However, relationships between two entities in a document can be expressed over more than one sentence. For example, *"Allied Capital announced interim results and a new CEO last week. John Scheurer has been appointed to the post with immediate effect."* also asserts a relation between *"John Scheurer"* and *"Allied Capital"*. We call these relations inter-sentential.

Previous work investigating the proportion of inter-sentential relations in information extraction corpora is limited. An examination of relations in single and multiple sentences within the MUC4, MUC6 and MUC7 corpora (Stevenson, 2007) estimated that at most 40% of events containing 2 or more entities were inter-sentential. An analysis carried out on a manually annotated corpus of clinical records of cancer patients (Roberts et al., 2008) reported that 22.7% of relations were inter-sentential.

It is important to know the proportion and distribution of inter-sentential relations in IE corpora both to understand the coverage of existing relation extraction systems and to inform new developments in relation extraction. This paper reports an investigation into the inter-sentential relations found in two widely used Information Extraction corpora: the Message Understanding Conference (MUC6) corpus from 1995 and the Automatic Content Extraction (ACE03) corpus from 2003. The analysis in this paper focuses on finding the distribution of inter-sentential and single-sentential relations in the MUC6 corpus and the ACE corpus from 2003.

In Section 2. we introduce the corpora used in this work. In Section 3. we describe the supplementary annotation carried out on the MUC6 corpus to locate entities participating in events within the documents. Section 4. and 5. detail the results of our analysis of the inter-sentential relations in the MUC6 and ACE03 corpora respectively.

2. Information Extraction Corpora

2.1. MUC6

The Message Understanding Conferences (MUC) were organised by US military research institutions as a mechanism for stimulating the research and development of information extraction systems. This study analyses the corpus sixth message understanding conference (MUC6) (Grishman and Sundheim, 1996b). The MUC6 scenario was to extract information about management succession events from newswire. The organisers provided a training corpus and a set of templates which contain the information about the management succession contained in the corpus. These templates include information such as the names of people who are starting or leaving management posts, the names of the respective posts and organisations, the reason for the vacancy and whether the named person is currently in the job. The templates were manually created by annotators who read each document. An inter-annotator agreement score of 83% was reported for this task. The information entered in the template was one of three types: a string taken directly from the document (e.g. a person's name, *"John Philips"*), an option from a list of predetermined fillers (e.g. is the

person joining or leaving a company, "IN" or "OUT") or a description by the annotator (e.g. comments). For example, the following sentences describe the event where Vern Raburn becomes president of the Paul Allen Group.

> "Paul G. Allen, the billionaire co-founder of Microsoft Corp., has started a company and named longtime friend **Vern Raburn** its **president** and chief executive officer.
> The company, to be called **Paul Allen Group**, will be based in **Bellevue**, Wash., and will..."

This event is encoded in the template shown in Figure 1[1]. The text shown in **bold font** in the above sentences become the string fillers for, respectively, the PER_NAME, POST, ORG_NAME and LOCALE fields in the event template in Figure 1. Additional information derived from these sentences are recorded in the list filler fields TYPE, NEW_STATUS, ON_THE_JOB and VACANCY_REASON. The key pieces of information that uniquely identify any management succession event in the corpus are: ORG_NAME, POST and PER_NAME and this study concentrates on relations between these entities.

2.2. ACE 2003

The Automatic Content Extraction (ACE) project is the successor to MUC and continues to create annotated corpora for relation extraction research. The ACE extraction tasks differ from those used in MUC in that ACE aims to extract domain independent binary relations from documents. Extracting pairs of entities that are linked by some semantic relationship is more straightforward than extracting multiple entities and deriving information to fill complex template structures. ACE has also moved away from highly domain specific extraction by annotating five generic relation types with 24 subtypes to further characterize them:

ROLE The role a person plays in an organisation; subtyped as Management, General-Staff, Member, Owner, Founder, Client, Affiliate-Partner, Citizen-Of, or Other.

PART Part-whole relationships; subtyped as Subsidiary, Part-Of, or Other.

AT Location relationships; subtyped as Located, Based-In, or Residence.

NEAR Identifying relative locations.

SOCIAL Personal and professional relationships between people; subtyped as Parent, Sibling, Spouse, Grandparent, Other-Relative, Other-Personal, Associate, or Other-Professional.

These relationships are of the kind that can be found in any news related text. An example of an AT relation, of subtype LOCATED, is expressed in the following sentences and produces the template in Table 2[2].

> "In **Moscow**, today there was a dramatic plot twist in a real life courtroom drama.
>
> **A lawyer for US spy suspect, Edwin Prope** said that the main prosecution witness has retracted his testimony."

The relations and entities are manually extracted by annotators. An inter-annotator agreement of 52% for the English relation extraction task was reported.
Since ACE04 the binary relation extraction task has been limited to annotating relations which are expressed within a single sentence. Consequently, we have used the ACE03 corpus in this study, the last ACE corpus to have inter-sentential relations annotated.

3. MUC6 Annotation

Determining the proportion of inter-sentential relations in each corpus requires that we count the number of relations that cross sentence boundaries. A mapping between the entity mentions and the documents is provided in the ACE03 annotations in the form of character offsets for the pair of entities participating in the relation. However, the MUC event templates do not specify where the entities in the templates are located in the documents. This section describes an additional annotation of the MUC6 corpus in which the strings describing each event in the templates are located within each document[3].
Following Stevenson (2007) the person(PER), organisation (ORG) and post (POST) slots in the template are chosen to represent the key information in the MUC6 management succession scenario. We define binary relations for MUC6 as a management succession relationship between any two of these entities (i.e. PER_ORG, PER_POST or ORG_POST). Multiple mentions of the PER, ORG or POST strings in the document leads to an ambiguity over which mention of the entity is actually participating in the management succession event. Unfortunately, this is often the case since many MUC6 news-stories focus on a single succession event, company or person. To ensure that this ambiguity does not lead to entity pairs being wrongly identified as relations in our analysis, every PER, ORG or POST string from the event template which occurs in a document *and* participates in a management succession event is manually identified in the document.
Two hundred documents were manually annotated using the corresponding event templates. An example of this annotation is shown in Figure 3. The three XML tags used for the annotation correspond to the entity types PER, ORG and POST. The common attribute ev is the event ID given in the template and whether the person is joining (IN) or leaving (OUT) the organisation is given directly after the event ID. The annotations in this example indicate that the person *"Richard C. Bartlett"* is joining an organisation called *"Mary Kay Corp."* in the post of *"vice-chairman"*. A

[1] This event template has been summarized for illustrative purposes. A full description of the MUC6 template structure is detailed in Grishman and Sundheim (1996a)

[2] This relation annotation has been summarized for illustrative purposes. A full description of the ACE corpus annotations are available from http://projects.ldc.upenn.edu/ace/annotation/

[3] This set of annotations is available from the authors.

SUCCESSION_EVENT_ID	9404150071-1	
ORGANIZATION	**NAME**	"Paul Allen Group"
	TYPE	COMPANY
	LOCALE	Bellevue CITY
PER_NAME	"Vern Raburn"	
NEW_STATUS	IN	
ON_THE_JOB	NO	
POST	"president"	
VACANCY_REASON	NEW_POST_CREATED	

Figure 1: Summarized event template for succession event 9404150071-1 in MUC6 corpus.

RELATION ID	PRI20001 108.2000.1506-R3	
TYPE	AT	
SUBTYPE	LOCATED	
CLASS	EXPLICIT	
ARGNUM1	**STRING**	"Moscow"
	START	69
	END	74
	TYPE	GPE
ARGNUM2	**STRING**	"A lawyer for US spy suspect, Edwin Prope"
	START	148
	END	187
	TYPE	PER

Figure 2: Summarized event template for binary relation PRI20001 108.2000.1506-R3 in the ACE corpus (2003).

```
<PER ev=1(IN)>Richard C. Bartlett</PER>
was named to the newly created position of
<POST ev=1(IN)>vice chairman</POST> of <ORG
ev=1(IN)>Mary Kay Corp.</ORG>, a privately
held cosmetics company.
```

Figure 3: MUC6 annotation

sample of 5 documents were annotated twice and the inter-annotator agreement between these found to be 84%.

All of the disagreements between the two annotators related to the markup of organsation names. This often occurred because one annotator chose to mark up the organisation mention most closely preceding the mentions of the other participating entities, deciding that the mention of the entity at that point in the discourse meant that it was implicitly participating in the succession event. Whereas the other annotator marked up the organisation mention that made some reference to the management succession event and only if this was not available reverted to marking up the closest preceding mention. For example, the following following sentences describe the replacement of Rosso by Wareham as president of Beckman Instruments Inc.

> "**Beckman Instruments Inc.** said it will cut 11% of its work force, consolidate its life sciences laboratory and diagnostic laboratory businesses, and take other steps to reorganize and cut costs.
>
> **Beckman's** stock dropped 50 cents to $26.25 a share in composite New York Stock Exchange trading.
>
> Louis Rosso, chairman and chief executive officer, said **John P. Wareham**, 52 years old, will succeed him as **president**."

One annotator marked-up the company name in the first sentence as it is mentioned in the context of reorganisation and cost cutting. While the other chose the company name in the second sentence which is the mention of the organisation name which most closely precedes person name and post participating in the event. The main purpose of our analysis is to determine whether events are inter-sentential and the disagreements between annotators did not alter the classification of any event (all related to inter-sentential events). We chose the first approach for the final annotation of the 200 documents.

4. MUC6 Analysis

Determining the number of relations which are asserted within a single sentence was carried out in two steps. First, the sentence boundaries in each document are found using the ANNIE sentence splitter, made available through the GATE framework (Cunningham et al., 2002). Next, the entities in each relation are located in each document using the relation annotations (see Section 3.). The resulting sentence boundary information and entity locations were used to count the number of inter-sentential relations (INTER) and single-sentence relations (SINGLE). The results are detailed in Table 1. The corpus of 200 documents contain a total of 1599 binary relations. The distribution of

	INTER	SINGLE	Total
PER_ORG	212 (39.2%)	329 (60.8%)	541
PER_POST	66 (12.4%)	465 (87.6%)	531
POST_ORG	177 (33.6%)	350 (66.4%)	527
Total	455 (28.5%)	1144 (71.5%)	1599

Table 1: MUC6 binary relation distribution

the different relation types shows that PER_POST relations are rarely inter-sentential (12.4%) while the proportion of inter-sentential PER_ORG relations is over three times this figure (39.2%). One reason for the greater proportion of inter-sentential PER_ORG and POST_ORG relations is because the news articles in this corpus often focus on a single company. The name of the company is mentioned at the beginning of the document and the referred to using anaphoric expressions in the remainder of the document. The lower proportion of inter-sentential POST_ORG to PER_ORG relations can be attributed to a common prepositional phrase construction used to express this relation, such as *"president of Merck & Co."*, *"chief executive officer of rival drug maker Eli Lilly & Co."* and *"CEO of Canadian telephone-equipment maker Northern Telecom Ltd."*

Through manual examination of PER_POST relations we observe two common linguistic constructs. The first is a subject-verb-object structure; for example, *"PER selected as POST"*, *"PER will resign as POST"* and *"PER will assume his post as POST"*. The second construct is a non-restrictive appositive; for example, *"...fired it's POST, PER, last week..."* and *"PER, the firm's POST, said..."*. In cases where a PER_POST relation is inter-sentential it is often expressed in consecutive sentences where the first sentence describes an incoming management succession event and the second sentence provides the complementing outgoing management succession event without restating information about the company or post. The sentence often uses the anaphoric construction of the kind *"He is succeeded by..."* or *"She will succeed..."*.

28.5% of the total number of relations in the MUC6 corpus are inter-sentential. This shows that a significant proportion of relations could not be captured by a single sentence extraction system.

Distribution of inter-sentential relations A further analysis carried out on the inter-sentential relations in the corpus determined the number of sentence boundaries between the two entities in each of the inter-sentential relations. The results are detailed in Table 2. We see that 46.6% of inter-sentential relations are contained within adjacent sentences and 66.6% are contained within a window of 3 sentences. This shows that two-thirds of inter-sentential relations can be found when limiting the search to a window of three sentences and that sentences further apart from each other are less likely to contain inter-sentential relations.

5. ACE03 Analysis

Unlike the MUC templates, the annotations for the ACE03 corpus include the location of the entities in the documents. Therefore, the only preprocessing the data required was sentence splitting which was also carried out using the ANNIE sentence splitter. The sentence boundaries together with the ACE annotations was used to determine which relations are inter-sentential. The results presented in Table 3 show the relation distribution over the 5 main relation types. The corpus contains 146 documents and a total of 1638 binary relations. 9.4% of these relations were found to be inter-sentential. The proportion of inter-sentential relations varies with relationship type. For example, the relation type NEAR has the highest proportion of inter-sentential relations at 21.9% whilst the PART relation is almost never expressed across sentences with 4.2% inter-sentential.

	INTER-SENTENTIAL			
	PER_ORG	PER_POST	POST_ORG	TOTAL
1	94	49	69	212
2	45	8	38	91
3	21	3	20	44
4	15	0	14	29
5	9	0	10	19
6	7	1	6	14
>6	21	5	20	46

Table 2: MUC6 distribution of inter-sentential relations across sentence boundaries.

An example of a inter-sentential AT relation is given below. This example shows a location relation between *"82 people"* and *"the runway"* which can only be understood by the reader using real-world knowledge. That is, we know that if a plane has exploded during take-off *and* 82 people were killed in the incident *then* those people must have been located on the runway. This example demonstrates that there exists a subset of inter-sentential relations which cannot automatically extracted using only syntax and reference.

> investigators say the boeing 747 hit a barrier and a crane on **the runway** and exploded during take-off.
>
> **82 people** were killed.

The ACE03 corpus has a far smaller, but still significant, proportion of inter-sentential relations compared with MUC6 that cannot be extracted using single-sentence extraction systems.

	INTER	SINGLE	Total
ROLE	77 (9.8%)	709 (90.2%)	786
AT	55 (11.1%)	440 (88.9%)	495
NEAR	7 (21.9%)	25 (78.1%)	32
PART	7 (4.2%)	161 (95.8%)	168
SOC	8 (5.2%)	147 (94.8%)	155
Total	154 (9.4%)	1482 (90.6%)	1636

Table 3: ACE03 binary relation distribution

6. Conclusion

This paper details an analysis of relation distribution in MUC6 and ACE03 corpora. We found that 28.5% of

MUC6 relations and 9.4% of ACE03 relations are inter-sentential. This quantification of inter-sentential relations in the MUC6 corpus (28.5%) improves on previous investigation (Stevenson, 2007) that calculated an upper bound of 40%.

The difference between the proportion of inter-sentential relations in the two corpora is due to the nature of the extraction task used by each. The MUC6 extraction task requires the filling of complex templates that are specific to the types of information contained in the documents that form the MUC6 corpus. Information from a variety of sentences in a document is often required to complete these templates. The MUC6 templates were simplified to a set of three binary relations for the study presented here but often still include information from multiple sentences. On the other hand the ACE03 task involves the extraction of simpler generic relations that may be found in a range of documents. These relations can be identified without considering the wider context of the document in a larger proportion of cases. It is therefore important to consider the nature of the extraction task when deciding whether to use a relation extraction system that is limited to identifying single-sentence relations.

This study has also shown that some relation types are more commonly asserted using cross-sentence linguistic constructs and that in the MUC6 corpus likelihood of two entities being related is inversely proportional to the number of sentence boundaries between the two entities. This information may be useful for the development of approaches to the detection of inter-sentential relations.

7. References

Aron Culotta and Jeffrey Sorensen. 2004. Dependency tree kernels for relation extraction. In *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 423, Morristown, NJ, USA. Association for Computational Linguistics.

Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. 2002. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*.

Ralph Grishman and Beth Sundheim. 1996a. Design of the MUC-6 evaluation. In *Proceedings of a workshop on held at Vienna, Virginia*, pages 413–422, Morristown, NJ, USA. Association for Computational Linguistics.

Ralph Grishman and Beth Sundheim. 1996b. Message Understanding Conference-6: a brief history. In *Proceedings of the 16th conference on Computational linguistics*, pages 466–471, Morristown, NJ, USA. Association for Computational Linguistics.

Angus Roberts, Robert Gaizauskas, and Mark Hepple. 2008. Extracting clinical relationships from patient narratives. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pages 10–18, Columbus, Ohio, June. Association for Computational Linguistics.

Mark Stevenson. 2007. Fact distribution in information extraction. *Language Resources and Evaluation*, 40(2):183–201, May.

Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2003. Kernel methods for relation extraction. *J. Mach. Learn. Res.*, 3:1083–1106.

Shubin Zhao and Ralph Grishman. 2005. Extracting relations with integrated information using kernel methods. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 419–426, Morristown, NJ, USA. Association for Computational Linguistics.

Guodong Zhou, Min Zhang, Dong Hong Ji, and QiaoMing Zhu. 2007. Tree kernel-based relation extraction with context-sensitive structured parse tree information. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 728–736, Prague, Czech Republic.