Information Extraction from Single and Multiple Sentences

Mark Stevenson
Department of Computer Science
Regent Court, 211 Portobello Street,
University of Sheffield
Sheffield
S1 4DP, UK
marks@dcs.shef.ac.uk

Abstract

Some Information Extraction (IE) systems are limited to extracting events expressed in a single sentence. It is not clear what effect this has on the difficulty of the extraction task. This paper addresses the problem by comparing a corpus which has been annotated using two separate schemes: one which lists all events described in the text and another listing only those expressed within a single sentence. It was found that only 40.6% of the events in the first annotation scheme were fully contained in the second.

1 Introduction

Information Extraction (IE) is the process of identifying specific pieces of information in text, for example, the movements of company executives or the victims of terrorist attacks. IE is a complex task and a the description of an event may be spread across several sentences or paragraphs of a text. For example, Figure 1 shows two sentences from a text describing management succession events (i.e. changes in corporate executive management personnel). It can be seen that the fact that the executives are leaving and the name of the organisation are listed in the first sentence. However, the names of the executives and their posts are listed in the second sentence although it does not mention the fact that the executives are leaving these posts. The succession events can only be fully understood from a combination of the information contained in both sentences.

Combining the required information across sentences is not a simple task since it is necessary to identify phrases which refer to the same entities, "two top executives" and "the executives" in the above example. Additional difficulties occur because the same entity may be referred to by a different linguistic unit. For example, "International Business Machines Ltd." may be referred to by an abbreviation ("IBM"), nickname ("Big Blue") or anaphoric expression such as "it" or "the company". These complications make it difficult to identify the correspondences between different portions of the text describing an event.

> Pace American Group Inc. said it notified two top executives it intends to dismiss them because an internal investigation found evidence of "self-dealing" and "undisclosed financial relationships." The executives are Don H. Pace, cofounder, president and chief executive officer; and Greg S. Kaplan, senior vice president and chief financial officer.

Figure 1: Event descriptions spread across two sentences

Traditionally IE systems have consisted of several components with some being responsible for carrying out the analysis of individual sentences and other modules which combine the events they discover. These systems were often designed for a specific extraction task and could only be modified by experts. In an effort to overcome this brittleness machine learning methods have been applied to port systems to new domains and extraction tasks with minimal manual intervention. However, some IE systems using machine learning techniques only extract events which are described within a single sentence, examples include (Soderland, 1999; Chieu and Ng, 2002; Zelenko et al., 2003). Presumably an assumption behind these approaches is that many of the events described in the text are expressed within a single sentence and there is little to be gained from the extra processing required to combine event descriptions.

Systems which only attempt to extract events described within a single sentence only report results across those events. But the proportion of events described within a single sentence is not known and this has made it difficult to compare the performance of those systems against ones which extract all events from text. This question is addressed here by comparing two versions of the same IE data set, the evaluation corpus used in the Sixth Message Understanding Conference (MUC-6) (MUC, 1995). The corpus produced for this exercise was annotated with all events in the corpus, including those described across multiple sentences. An independent annotation of the same texts was carried out by Soderland (1999), although he only identified events which were expressed within a single sentence. Directly comparing these data sets allows us to determine what proportion of all the events in the corpus are described within a single sentence.

The remainder of this paper is organised as follows. Section 2 describes the formats for representing events used in the MUC and Soderland data sets. Section 3 introduces a common representation scheme which allows events to be compared, a method for classifying types of event matches and a procedure for comparing the two data sets. The results and implications of this experiment are presented in Section 4. Some related work is discussed in Section 5.

2 Event Scope and Representation

The topic of the sixth MUC (MUC-6) was management succession events (Grishman and Sundheim, 1996). The MUC-6 data has been commonly used to evaluate IE systems. The test corpus consists of 100 *Wall Street Journal* documents from the period January 1993 to June 1994, 54 of which contained management succession events (Sundheim, 1995). The format used to represent events in the MUC-6 corpus is now described.

2.1 MUC Representation

Events in the MUC-6 evaluation data are recorded in a nested template structure. This format is useful for representing complex events which have more than one participant, for example, when one executive leaves a post to be replaced by another. Figure 2 is a simplified event from the the MUC-6 evaluation similar to one described by Grishman and Sundheim (1996).

```
<SUCCESSION_EVENT-9402240133-2> :=
    SUCCESSION_ORG:
        <ORGANIZATION-9402240133-1>
    POST: "chairman"
    IN_AND_OUT: <IN_AND_OUT-9402240133-4>
    VACANCY_REASON: DEPART_WORKFORCE

<IN_AND_OUT-9402240133-4> :=
    IO_PERSON: <PERSON-9402240133-1>
    NEW_STATUS: IN
    ON_THE_JOB: NO
    OTHER_ORG: <ORGANIZATION-9402240133-1>
    REL_OTHER_ORG: SAME_ORG

<ORGANIZATION-9402240133-1> :=
    ORG_NAME: "McCann-Erickson"
    ORG_ALIAS: "McCann"
    ORG_TYPE: COMPANY

<PERSON-9402240133-1> :=
    PER_NAME: "John J. Dooner Jr."
    PER_ALIAS: "John Dooner"
               "Dooner"
```

Figure 2: Example Succession event in MUC format

This template describes an event in which “John J. Dooner Jr.” becomes chairman of the company “McCann-Erickson”. The MUC templates are too complex to be described fully here but some relevant features can be discussed. Each SUCCESSION_EVENT contains the name of the POST, organisation (SUCCESSION_ORG) and references to at least one IN_AND_OUT sub-template, each of which records an event in which a person starts or leaves a job. The IN_AND_OUT sub-template contains details of the PERSON and the NEW_STATUS field which records whether the person is starting a new job or leaving an old one.

Several of the fields, including POST, PERSON and ORGANIZATION, may contain aliases which are alternative descriptions of the field filler and are listed when the relevant entity was described in different was in the text. For example, the organisation in the above template has two descriptions: “McCann-Erickson” and “McCann”. It should be noted that the MUC template structure does not link the field fillers onto particular instances in the texts. Consequently if the same entity description is used more than once then there is no simple way of identifying which instance corresponds to the event description.

The MUC templates were manually filled by annotators who read the texts and identified the management succession events they contained. The MUC organisers provided strict guidelines about what constituted a succession event and how the templates should be filled which the annotators sometimes found difficult to interpret (Sundheim, 1995). Interannotator agreement was measured on 30 texts which were examined by two annotators. It was found to be 83% when one annotator's templates were assumed to be correct and compared with the other.

2.2 Soderland's Representation

Soderland (1999) describes a supervised learning system called WHISK which learned IE rules from text with associated templates. WHISK was evaluated on the same texts from the MUC-6 data but the nested template structure proved too complex for the system to learn. Consequently Soderland produced his own simpler structure to represent events which he described as "case frames". This representation could only be used to annotate events described within a single sentence and this reduced the complexity of the IE rules which had to be learned.

The succession event from the sentence "Daniel Glass was named president and chief executive officer of EMI Records Group, a unit of London's Thorn EMI PLC." would be represented as follows:[1]

```
@@TAGS Succession
{PersonIn DANIEL GLASS}
{Post PRESIDENT AND CHIEF EXECUTIVE OFFICER}
{Org EMI RECORDS GROUP}
```

Events in this format consist of up to four components: PersonIn, PersonOut, Post and Org. An event may contain all four components although none are compulsory. The minimum possible set of components which can form an event are (1) PersonIn, (2) PersonOut or (3) both Post and Org. Therefore a sentence must contain a certain amount of information to be listed as an event in this data set: the name of an organisation and post participating in a management succession event or the name of a person changing position and the direction of that change.

Soderland created this data from the MUC-6 evaluation texts without using any of the existing annotations. The texts were first pre-processing using the University of Massachusetts BADGER syntactic analyser (Fisher et al., 1995) to identify syntactic clauses and the named entities relevant to the management succession task: people, posts and organisations. Each sentence containing relevant entities was examined and succession events manually identified.

This format is more practical for machine learning research since the entities which participate in the event are marked directly in the text. The learning task is simplified by the fact that the information which describes the event is contained within a single sentence and so the feature space used by a learning algorithm can be safely limited to items within that context.

3 Event Comparison

3.1 Common Representation and Transformation

There are advantages and disadvantages to the event representation schemes used by MUC and Soderland. The MUC templates encode more information about the events than Soderland's representation but the nested template structure can make them difficult to interpret manually.

In order to allow comparison between events each data set was transformed into a common format which contains the information stored in both representations. In this format each event is represented as a single database record with four fields: type, person, post and organisation. The type field can take the values person_in, person_out or, when the direction of the succession event is not known, person_move. The remaining fields take the person, position and organisation names from the text. These fields may contain alternative values which are separated by a vertical bar ("|").

MUC events can be translated into this format in a straightforward way since each IN_AND_OUT sub-template corresponds to a single event in the common representation. The MUC representation is more detailed than the one used by Soderland and so some information is discarded from the MUC templates. For example, the VACANCY_REASON filed which lists the reason for the management succession event is not transfered to the common format. The event listed in Figure 2 would be represented as follows:

```
type(person_in)
person(‘John J. Dooner Jr.’|
       ‘John Dooner’|‘Dooner’)
org(‘McCann-Erickson’|‘McCann’)
post(chairman)
```

Alternative fillers for the person and org fields are listed here and these correspond to the PER_NAME, PER_ALIAS, ORG_NAME and ORG_ALIAS

[1] The representation has been simplified slightly for clarity.

fields in the MUC template.

The Soderland succession event shown in Section 2.2 would be represented as follows in the common format.

```
type(person_in)
person(‘Daniel Glass’)
post(‘president’)
org(‘EMI Records Group’)

type(person_in)
person(‘Daniel Glass’)
post(‘chief executive officer’)
org(‘EMI Records Group’)
```

In order to carry out this transformation an event has to be generated for each `PersonIn` and `PersonOut` mentioned in the Soderland event. Soderland's format also lists conjunctions of post names as a single slot filler (“president and chief executive officer” in this example). These are treated as separate events in the MUC format. Consequently they are split into the separate post names and an event generated for each in the common representation.

It is possible for a Soderland event to consist of only a `Post` and `Org` slot (i.e. there is neither a `PersonIn` or `PersonOut` slot). In these cases an underspecified type, `person_move`, is used and no `person` field listed. Unlike MUC templates Soderland's format does not contain alternative names for field fillers and so these never occur when an event in Soderland's format is translated into the common format.

3.2 Matching

The MUC and Soderland data sets can be compared to determine how many of the events in the former are also contained in the latter. This provides an indication of the proportion of events in the MUC-6 domain which are expressible within a single sentence. Matches between Soderland and MUC events can be classified as **full**, **partial** or **nomatch**. Each of these possibilities may be described as follows:

Full A pair of events can only be fully matching if they contain the same set of fields. In addition there must be a common filler for each field. The following pair of events are an example of two which fully match.

```
type(person_in)
person(‘R. Wayne Diesel’|‘Diesel’)
org(‘Mechanical Technology Inc.’|
    ‘Mechanical Technology’)
post(‘chief executive officer’)

type(person_in)
person(‘R. Wayne Diesel’)
org(‘Mechanical Technology’)
post(‘chief executive officer’)
```

Partial A partial match occurs when one event contains a proper subset of the fields of another event. Each field shared by the two events must also share at least one filler. The following event would partially match either of the above events; the `org` field is absent therefore the matches would not be full.

```
type(person_in)
person(‘R. Wayne Diesel’)
post(‘chief executive officer’)
```

Nomatch A pair of events do not match if the conditions for a full or partial match are not met. This can occur if corresponding fields do not share a filler or if the set of fields in the two events are not equivalent or one the subset of the other.

Matching between the two sets of events is carried out by going through each MUC event and comparing it with each Soderland event for the same document. The MUC event is first compared with each of the Soderland events to check whether there are any equal matches. If one is found a note is made and the matching process moves onto the next event in the MUC set. If an equal match is not found the MUC event is again compared with the same set of Soderland events to see whether there are any partial matches. We allow more than one Soderland event to partially match a MUC event so when one is found the matching process continues through the remainder of the Soderland events to check for further partial matches.

4 Results

4.1 Event level analysis

After transforming each data set into the common format it was found that there were 276 events listed in the MUC data and 248 in the Soderland set. Table 1 shows the number of matches for each data set following the matching process described in Section 3.2. The counts under the "MUC data" and "Soderland data" headings list the number of events which fall into each category for the MUC and Soderland data sets respectively along with corresponding percentages of that data set. It can be seen that 112 (40.6%) of the MUC events are fully covered by the second data set, and 108 (39.1%) partially covered.

Match	MUC data		Soderland data	
Type	Count	%	Count	%
Full	112	40.6%	112	45.2%
Partial	108	39.1%	118	47.6%
Nomatch	56	20.3%	18	7.3%
Total	276		248	

Table 1: Counts of matches between MUC and Soderland data.

Table 1 shows that there are 108 events in the MUC data set which partially match with the Soderland data but that 118 events in the Soderland data set record partial matches with the MUC data. This occurs because the matching process allows more than one Soderland event to be partially matched onto a single MUC event. Further analysis showed that the difference was caused by MUC events which were partially matched by two events in the Soderland data set. In each case one event contained details of the move type, person involved and post title and another contained the same information without the post title. This is caused by the style in which the newswire stories which make up the MUC corpus are written where the same event may be mentioned in more than one sentence but without the same level of detail. For example, one text contains the sentence "Mr. Diller, 50 years old, succeeds Joseph M. Segel, who has been named to the post of chairman emeritus." which is later followed by "At that time, it was announced that Diller was in talks with the company on becoming its chairman and chief executive upon Mr. Segel's scheduled retirement this month."

Table 1 also shows that there are 56 events in the MUC data which fall into the nomatch category. Each of these corresponds to an event in one data set with no corresponding event in the other. The majority of the unmatched MUC events were expressed in such a way that there was no corresponding event listed in the Soderland data. The events shown in Figure 1 are examples of this. As mentioned in Section 2.2, a sentence must contain a minimum amount of information to be marked as an event in Soderland's data set, either name of an organisation and post or the name of a person changing position and whether they are entering or leaving. In Figure 1 the first sentence lists the organisation and the fact that executives were leaving. The second sentence lists the names of the executives and their positions. Neither of these sentences contains enough information to be listed as an event under Soderland's representation, consequently the MUC events generated from these sentences fall into the nomatch category.

It was found that there were eighteen events in the Soderland data set which were not included in the MUC version. This is unexpected since the events in the Soderland corpus should be a subset of those in the MUC corpus. Analysis showed that half of these corresponded to spurious events in the Soderland set which could not be matched onto events in the text. Many of these were caused by problems with the BADGER syntactic analyser (Fisher et al., 1995) used to pre-process the texts before manual analysis stage in which the events were identified. Mistakes in this pre-processing sometimes caused the texts to read as though the sentence contained an event when it did not. We examined the MUC texts themselves to determine whether there was an event rather than relying on the pre-processed output.

Of the remaining nine events it was found that the majority (eight) of these corresponded to events in the text which were not listed in the MUC data set. These were not identified as events in the MUC data because of the the strict guidelines, for example that historical events and non-permanent management moves should not be annotated. Examples of these event types include "... Jan Carlzon, who left last year after his plan for a merger with three other European airlines failed." and "Charles T. Young, chief financial officer, stepped down voluntarily on a 'temporary basis pending conclusion' of the investigation." The analysis also identified one event in the Soderland data which appeared to correspond to an event in the text but was not listed in the MUC scenario template for that document. It could be argued that there nine events should be added to the set of MUC events and treated as fully matches. However, the MUC corpus is commonly used as a gold standard in IE evaluation and it was decided not to alter it. Analysis indicated that one of these nine events would have been a full match and eight partial matches.

It is worth commenting that the analysis carried out here found errors in both data sets. There appeared to be more of these in the Soderland data but this may be because the event structures are much easier to interpret and so errors can be more readily identified. It is also difficult to interpret the MUC guidelines in some cases and it sometimes necessary to make a judgement over how they apply to a particular event.

4.2 Event Field Analysis

A more detailed analysis can be carried out examining the matches between each of the four fields in the event representation individually. There are 1,094 fields in the MUC data. Although there are 276 events in that data set seven of them do not mention a post and three omit the organisation name. (Organisation names are omitted from the template when the text mentions an organisation description rather than its name.)

Table 4.2 lists the number of matches for each of the four event fields across the two data sets. Each of the pairs of numbers in the main body of the table refers to the number of matching instances of the relevant field and the total number of instances in the MUC data.

The column headed "Full match" lists the MUC events which were fully matched against the Soderland data and, as would be expected, all fields are matched. The column marked "Partial match" lists the MUC events which are matched onto Soderland fields via partially matching events. The column headed "No-match" lists the event fields for the 56 MUC events which are not represented at all in the Soderland data.

Of the total 1,094 event fields in the MUC data 727, 66.5%, can be found in the Soderland data. The rightmost column lists the percentages of each field for which there was a match. The counts for the type and person fields are the same since the type and person fields are combined in Soderland's event representation and hence can only occur together. These figures also show that there is a wide variation between the proportion of matches for the different fields with 76.8% of the person and type fields being matched but only 43.2% of the organisation field.

This difference between fields can be explained by looking at the style in which the texts forming the MUC evaluation corpus are written. It is very common for a text to introduce a management succession event near the start of the newswire story and this event almost invariably contains all four event fields. For example, one story starts with the following sentence: "Washington Post Co. said Katharine Graham stepped down after 20 years as chairman, and will be succeeded by her son, Donald E. Graham, the company's chief executive officer." Later in the story further succession events may be mentioned but many of these use an anaphoric expression (e.g. "the company") rather than explicitly mention the name of the organisation in the event. For example, this sentence appears later in the same story: "Alan G. Spoon, 42, will succeed Mr. Graham as president of the company." Other stories again may only mention the name of the person in the succession event. For example, "Mr. Jones is succeeded by Mr. Green" and this explains why some of the organisation fields are also absent from the partially matched events.

4.3 Discussion

From some perspectives it is difficult to see why there is such a difference between the amount of events which are listed when the entire text is viewed compared with considering single sentences. After all a text comprises of an ordered list of sentences and all of the information the text contains must be in these. Although, as we have seen, it is possible for individual sentences to contain information which is difficult to connect with the rest of the event description when a sentence is considered in isolation.

The results presented here are, to some extent, dependent on the choices made when representing events in the two data sets. The events listed in Soderland's data require a minimal amount of information to be contained within a sentence for it to be marked as containing information about a management succession event. Although it is difficult to see how any less information could be viewed as representing even part of a management succession event.

5 Related Work

Huttunen et al. (2002) found that there is variation between the complexity of IE tasks depending upon how the event descriptions are spread through the text and the ways in which they are encoded linguistically. The analysis presented here is consistent with their finding as it has

	Full match	Partial match	Nomatch	TOTAL	%
Type	112 / 112	100 / 108	0 / 56	212 / 276	76.8%
Person	112 / 112	100 / 108	0 / 56	212 / 276	76.8%
Org	112 / 112	6 / 108	0 / 53	118 / 273	43.2%
Post	111 / 111	74 / 108	0 / 50	185 / 269	68.8%
Total	447 / 447	280 / 432	0 / 215	727 / 1094	66.5%

Table 2: Matches between MUC and Soderland data at field level

been observed that the MUC texts are often written in such as way that the name of the organisation in the event is in a different part of the text to the rest of the organisation description and the entire event can only be constructed by resolving anaphoric expressions in the text. The choice over which information about events should be extracted could have an effect on the difficulty of the IE task.

6 Conclusions

It seems that the majority of events are not fully described within a single sentence, at least for one of the most commonly used IE evaluation sets. Only around 40% of events in the original MUC data set were fully expressed within the Soderland data set. It was also found that there is a wide variation between different event fields and some information may be more difficult to extract from text when the possibility of events being described across multiple sentences is not considered. This observation should be borne in mind when deciding which approach to use for a particular IE task and should be used to put the results reported for IE systems which extract from a single sentence into context.

Acknowledgements

I am grateful to Stephen Soderland for allowing access to his version of the MUC-6 corpus and advice on its construction. Robert Gaizauskas and Beth Sundheim also provided advice on the data used in the MUC evaluation. Mark Hepple provided valuable comments on early drafts of this paper. I am also grateful to an anonymous reviewer who provided several useful suggestions.

References

H. Chieu and H. Ng. 2002. A Maximum Entroy Approach to Information Extraction from Semi-structured and Free Text. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence (AAAI-02)*, pages 768–791, Edmonton, Canada.

D. Fisher, S. Soderland, J. McCarthy, F. Feng, and W. Lehnert. 1995. Description of the UMass system as used for MUC-6. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 221–236, San Francisco, CA.

R. Grishman and B. Sundheim. 1996. Message understanding conference - 6 : A brief history. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96)*, pages 466–470, Copenhagen, Denmark.

S. Huttunen, R. Yangarber, and R. Grishman. 2002. Complexity of Event Structures in IE Scenarios. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING-2002)*, pages 376–382, Taipei, Taiwan.

MUC. 1995. *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, San Mateo, CA. Morgan Kaufmann.

S. Soderland. 1999. Learning Information Extraction Rules for Semi-structured and free text. *Machine Learning*, 31(1-3):233–272.

B. Sundheim. 1995. Overview of results of the MUC-6 evaluation. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 13–31, Columbia, MA.

D. Zelenko, C. Aone, and A. Richardella. 2003. Kernel methods for relation extraction. *Journal of Machine Learning Research*, 3:1083–1106.