Do coreferential arguments make event mentions coreferential?

Laura Hasler and Constantin Orăsan

Research Group in Computational Linguistics
University of Wolverhampton, UK
{L.Hasler, C.Orasan}@wlv.ac.uk

Abstract. This paper explores features of arguments in coreferential chains of events using the NP4E corpus, which has been previously annotated for entity and event coreference. The analysis takes event chains as a starting point and examines the way entities in the argument slots of event mentions contribute to establishing coreference. The investigation revealed that in many cases the arguments of coreferential event mentions are not coreferential themselves. The paper offers an explanation of this phenomenon and discusses the notion of event coreference with the help of examples from the corpus.

1 Introduction

Previous work in computational linguistics which deals with event coreference has either left the relations between the arguments of coreferential event mentions underspecified or has assumed that they can only be coreferential. The former is usually met in ACE-derived research, whilst the latter was common in early work on event coreference resolution, which was viewed as an extension of entity coreference [1]. However, an analysis of a corpus annotated for both entity and event coreference revealed that this assumption of coreference is not necessarily the case. Expectations of coreference-only relations need to be treated with caution as there are many cases where this constraint must be relaxed if all relevant event mentions are to be included in coreferential chains.

Events are notoriously difficult to define, and are often broadly characterised as things that happen, but this characterisation has been criticised for just shifting the task to one of clarifying the meaning of happen (Stanford Encyclopedia of Philosophy). However, for our purposes, and following related work such as the ACE program [2] and TimeBank/TimeML [3], [4], this is generally what we consider events to be: a thing that happens or takes place; a single specific occurrence, either instantaneous or ongoing, that is unique and can be anchored at a point in time. This definition was derived from related work in combination with a corpus analysis of news texts and applied in the annotation of events in the NP4E corpus [5]. The annotation did not only mark events and their arguments, i.e., participants and other attributes such as time and place, but also indicated which mentions refer to the same event and therefore constitute chains

1 http://plato.stanford.edu/entries/events/
of coreferential events. For the annotation of entity coreference in the same corpus, the established definition that two NPs are coreferential if they refer to the same entity in the real world [6] was used. This paper investigates features of event mentions marked as coreferential in the NP4E corpus, focusing on their arguments. The analysis takes coreferential chains of events as a starting point and examines the way entities in the argument slots of event mentions contribute to establishing coreference.

The rest of the paper is structured as follows: Section 2 briefly presents related work dealing with events. A description of the NP4E corpus and the annotation applied is provided in Section 3. Section 4 focuses on the analysis of coreferential event chains from the corpus. A comparison of the annotations by different annotators is given in Section 5. The paper finishes with conclusions and ideas for future work.

2 Related work

Because the work reported in this paper deals with the annotation of event mentions and coreferential chains of events, and through these relations between entities, related work can be found in all three areas. Most of this work looks at events from a different perspective, so whilst related, there are substantial differences between the NP4E annotation and analysis and the work described here.

In terms of the annotation of anaphora (including coreference), there are a number of established annotation schemes and corpora (e.g. [7], [8], [9]), but these do not tend to be linked event coreference. The ARRAU corpus [10] considers clauses as antecedents for discourse deixis which can be referred to by NPs, and includes certain bridging relations (part-of, set member, converse). However, the aim of the annotation in ARRAU corpus is not to annotate event arguments and therefore the annotation is approached differently. Work on event coreference started in the late 1990s [1], [11], and where event arguments were considered, they were expected to be coreferential.

OntoNotes [12] develops a multilingual, domain-independent resource to represent literal language incorporating predicate structure, word sense, ontology linking and coreference. In terms of coreference, the focus is on NPs although verb phrases may be marked when coreferential with an NP. In this sense the annotation is closer to that in the ARRAU corpus than the NP4E corpus. Related to the idea of addressing predicate structure is the development of resources which which deal with the annotation of semantic frames and arguments in various ways. FrameNet [13] consists of semantic frames describing situations and entities by annotating their arguments and roles to give a structured semantic representation of language. Whilst the focus of the NP4E project was on the annotation of coreferential entities and events, there are parallels in terms of annotating arguments for lexical items representing or evoking a wider situation/event and this resource could inform future annotations of similar phenomena.

The ACE program [2] has produced resources most closely related to our work, with corpora containing annotations of events, values, relations and entities for the evaluation of systems which recognise these phenomena for extraction. ACE 2005 [14], [15] annotates eight types of events, each with their own sub-types, including LIFE (INJURE, DIE), CONFLICT (ATTACK) and CONTACT and arguments, which are used in the NP4E annotation (see Section 3). Relations such as part-whole and organisation-affiliation.
and artifact are marked between entities, but it is not clear exactly how these relate to the event annotation as the relations can only be annotated between entities in the same sentence. Coreference is mentioned in the event annotation guidelines, but only briefly. Ji et al. [16] propose a new ACE-related task in cross-document event coreference which is based on events involving a particular person (or centroid argument). This takes the NP as the starting point rather than the events themselves in which that NP is involved and focuses on event extraction for IE.

Bejan and Harabagiu [17] describe a linguistic resource to annotate event structures and relations between events themselves rather than between arguments of the same chain. They approach event coreference from a TimeML angle, which was too broad for annotation in the NP4E corpus. TimeML [3] and its related resource TimeBank [4] are concerned with events but focus very much on the temporal aspect. This and the ACE-related work described above are geared towards automatic event processing and are not concerned with what actually happens when people try to annotate event coreference. From this point of view, the research presented in this paper is closer to the annotation of anaphora mentioned above.

3 Corpus description and annotation

The NP4E corpus contains clusters of texts in the domain of terrorism/security. The texts were collected in the context of a project which focused on designing annotation guidelines for NP and event coreference, and therefore the annotated corpus is of a limited size. The corpus contains five clusters of security-related newswire texts which report on bombings, hostage situations and hijacks, and totals approximately 55,000 words. The whole corpus was annotated for noun phrase coreference, but due to the difficulty of the task of event annotation, only two clusters were annotated for events and event coreference. The events section of the corpus contains just over 12,500 words and comprises 20 annotated files from the Bukavu bombing and Peru hostages clusters. Two annotators performed the annotation task, each annotating 5 texts (approximately 6,300 words) from each cluster. Two Bukavu and two Peru texts (2,840 words) were common to both annotators to allow a comparison of their annotations.

3.1 Annotation of events

The annotation was performed in two separate stages: NP annotation and event annotation. The first stage focused on marking coreferential chains of entities represented by noun phrases without considering that these NPs may also be involved in events. In the second stage, a specific set of events which best reflects those reported in the document clusters was annotated by marking event triggers and arguments. Event triggers are lexical units which best represent the event being reported. They can be lexicalised as verbs, nouns, adjectives and pronouns. Triggers representing five categories of events were annotated: ATTACK, DEFEND, INJURE, DIE, CONTACT. Because the texts are about bombings and a hostage crisis, the most prevalent categories

2 The terms trigger and arguments are taken from ACE.
of events were ATTACK and CONTACT. Annotators were instructed to mark coreference between triggers referring to exactly the same event.

Event arguments are participants and attributes related to the trigger that help to determine the event. Whilst participants are entities which take part or are involved in an event, attributes play a different role but are still related to an event and help to define it. Arguments depend on the category of the event under consideration, but the attributes TIME and PLACE are common to all event categories. For example, in an ATTACK event, an annotator can identify three participant slots: ATTACKER, MEANS and TARGET, and two attribute slots: TIME and PLACE. In a CONTACT event, one participant slot, AGENT (which can contain more than one entity) can be identified, along with three attribute slots: MEDIUM, TOPIC and TIME. The annotation guidelines fully specify which types of entity or expression can appear in different argument slots, for example:

- ATTACKER: PERSON, ORGANISATION, OTHER.
- MEANS: WEAPON, VEHICLE, OTHER.
- PLACE: LOCATION, FACILITY, OTHER.

The arguments identified during event annotation are NPs which were marked in the first stage. Given the complexity of the task, the annotation of arguments was limited to NPs within the same sentence as the trigger, or the surrounding sentences if this was necessary to disambiguate the event from other similar events or sub-events.\(^3\)

### 3.2 Annotation of event coreference

Rather than restrict the annotators by imposing on them that event triggers marked as coreferential must have arguments which also corefer when present in the same slots, the annotators were instructed to mark coreference between triggers which unambiguously refer to the same event occurring in the real world. To do this, they needed to consider the participants involved in the events as well as the time and location in each particular mention. However, they were not explicitly told to make sure that these aspects of the events also corefer. This permits us to examine whether it is necessary for all event arguments in the same slot across event mentions to corefer in order to consider these event mentions coreferential. Given that the event arguments are entities which may be part of coreferential chains, our initial hypothesis was that the arguments of coreferential triggers will also be coreferential when they appear in the same slot. This hypothesis is investigated in the next section.

### 4 Analysis and discussion

To perform the analysis, coreferential chains of events and their arguments were extracted from the corpus. Within the 344 unique event mentions, 106 coreferential chains, consisting of between 2 and 10 triggers and their arguments, were annotated. The remaining 238 events were referred to by only one trigger and were not considered coreferential with any other mentions. An example of an ATTACK event chain is

\(^3\) More details, including the annotation guidelines and the corpus can be found on the project’s website: http://clg.wlv.ac.uk/projects/NP4E
Zaire planes **bombs** rebels as U.N. seeks war’s end.

Zaire said on Monday its warplanes were **bombing** three key rebel-held towns in its eastern border provinces and that the raids would increase in intensity.

**Table 1.** Example of an ATTACK coreferential event chain

| TRIGGER: bombing |
| ATTKER: Zaire |
| MEANS: its warplanes |
| PLACE: three key rebel-held towns in its eastern border provinces |
| TIME: Monday |

"Since this morning the FAZ (Zaire army) has been **bombing** Bukavu, Shabunda and Walikale", said a defence ministry statement in the capital Kinshasa.

4.1 Referential relations between arguments in the same slot

An investigation of the relations between arguments of coreferential event mentions in the same slot showed that they often do not corefer. In a large number of cases the NPs assigned to the same argument slots are in an indirect anaphoric relationship such as part-of, set-member and subset. Of the 104 chains under consideration, 22 (21.15%) contained only coreferential NPs in their argument slots. The number of

---

4 Two of the 106 coreferential chains were not considered in this part of the analysis due to annotator errors.
chains comprising only non-coreferential arguments was 23 (22.12%). Ignoring the 9 chains which had too many empty slots for coreference to be observed (i.e. less than 2 fillers for the same slot), the remaining 50 chains contained a mixture of arguments which demonstrated both coreferential and indirect anaphoric relations. If one works under the assumption that coreferential event chains must contain arguments which also corefer, 70.19% of chains in the NP4E corpus would be affected: 22.12% would be ignored completely and 48.07% would have at least some elements missing.

In terms of argument slots rather than event chains, 200 slots with two or more fillers were identified across the corpus. Just 89 of these slots, less than half (44.5%), were filled with coreferential arguments only. None of the arguments corefer in 71 slots (35.5%) and 40 slots (20%) demonstrate a mixture of arguments in coreferential and non-coreferential relations. Again, if we restrict the fillers of argument slots to those in only a coreferential relationship with the other fillers, more than half are affected and a substantial number of slots would be emptied and elements removed from chains.

As an example, consider the ATTACK event presented in Table 1. The TARGET is both the rebel-held towns (three key rebel-held towns in its eastern border provinces and Bukavu, Shabunda and Walikale) and the rebels themselves who are in these towns (rebels). Whilst the towns and the rebels are not coreferential and can are only related because they are targets of the attack by being in the same place, they are both the targets of the same ATTACK event represented by different triggers in the text. Therefore they are both assigned to the TARGET slot and are in a part-of relationship. The same can be said for the attacker slot, which for one trigger has the argument Zaire and for the other the FAZ (Zaire army). In this case, there is a different kind of indirect relation between them, in that Zaire is used here as a political entity, which has an army working for it and carrying out the bombing on its behalf. Similarly, its warplanes is only a subset of Zaire planes. The temporal expressions used to fill the TIME slot are another example: this morning is a part of Monday (once Monday and today are established as the same day).

Where there are a combination of coreferential and non-coreferential arguments in a slot, there are often (where there are enough arguments) two or three groups of coreferential entities either in indirect anaphoric relationships with the other groups or also part of other event chains related to theirs. The example in Table 2 highlights one of these cases. There are 3 main NP chains: 20, 32 and 40. Chains 32 and 40, referring to Fujimori and Hashimoto are both in an indirect set-member relationship with chain 20 (the leaders of both nations, they). Chain 14, referring to rebels, is in a more distant relationship with the other slot fillers; one which is not clear from looking at this event chain alone. This relationship is permitted by being involved in the same event (the summit) which can take any number of agent arguments.

That is not to say that coreferential relations are not important. As stated above, there are a number of instances where chains have only this type of relation between their arguments. Coreferential arguments are particularly useful in establishing coreferential event mentions when the triggers which signal the events are different. However, if we are to consider using the relations between NP arguments in event

---

5 Most of the other slots in this chain were not filled, therefore we have shown only the relevant, filled AGENT slots.
<table>
<thead>
<tr>
<th>ID</th>
<th>TRIGGER</th>
<th>ARGUMENT: AGENT(S)</th>
</tr>
</thead>
<tbody>
<tr>
<td>c389</td>
<td>an emergency summit</td>
<td>the leaders of both nations: ID=20; CHAIN=20: PERSON</td>
</tr>
<tr>
<td>c397</td>
<td>the two-hour closed meeting</td>
<td>they: ID=24; CHAIN=20: PERSON</td>
</tr>
<tr>
<td>c408</td>
<td>the summit</td>
<td>Fujimori’s: ID=60; CHAIN=32: PERSON; Hashimoto: ID=58; CHAIN=40: PERSON</td>
</tr>
<tr>
<td>c409</td>
<td>the summit</td>
<td>Fujimori’s: ID=60; CHAIN=32: PERSON; Hashimoto: ID=58; CHAIN=40: PERSON</td>
</tr>
<tr>
<td>c418</td>
<td>the summit’s</td>
<td>rebels: ID=110; CHAIN=14: PERSON</td>
</tr>
<tr>
<td>c432</td>
<td>the summit</td>
<td>he: ID=170; CHAIN=40: PERSON</td>
</tr>
</tbody>
</table>

Table 2. Indirect anaphoric relations between AGENT slots in an event chain

coreference and if we want to include all relevant mentions in event chains, we need to take into account not only coreference, but also indirect referential relations. This is in keeping with the observation by Poesio and Artstein [10] in describing the ARRAU corpus that anaphora is not an equivalence relation and that markables form more complex structures than equivalence sets indicating identity of reference, particularly in the case of abstract entities and events.

Another case where the arguments of triggers in event chains are not coreferential is where there is indefinite quantification in a numerical value or an NP, such as up to 37 and at least six people (see Table 3). On several occasions in the corpus, even though the victim slot is filled twice with the same string these occurrences are not coreferential because they do not indicate identity of reference and we cannot be sure that exactly the same people are considered by each instance. Such examples are commonplace in news texts where the reporting of an event evolves and new information is added, or there are several parties reporting the event, which in some cases contradict each other or update previous information. There were several cases of this in the NP4E corpus. Other examples include two instances of at least nine people and one of nine confirmed dead, from the VICTIM slot in a DIE event chain, and hundreds, at least 500 people, the first group of at least 500 people, but probably more than that and the 500 annotated as both DEFENDER and ENTITY in a DEFEND event with the similar triggers flee, fled and left, and the PLACE slot filled with coreferential mentions of Bukavu.

Marking events containing such entities highlighted some issues in the annotation. Although the victims in those mentions may not corefer, the event of, for example, the plane dropping the bomb which injures people is still the same event. The INJURE event mentions might corefer but there is not enough evidence. Therefore, in the cases of DIE, INJURE, and possibly also DEFEND events, it may not desirable that the triggers are marked as coreferential when the VICTIM arguments are not coreferential. From an ATTACK point of view, the events would be coreferential, but because it is not clear exactly how many people were killed or injured when an indefinite quantifier is used, it is safest to say that these mentions are not coreferential.
at least nine people were killed and up to 37 wounded

i342 TRIGGER: wounding
AGENT: Zaire’s government: ID=5: CHAIN=5: ORGANISATION
VICTIM: up to 37: ID=16: CHAIN=16: PERSON
CAUSE: –
PLACE: one town: ID=17: CHAIN=17: LOCATION
TIME: Monday: ID=7: CHAIN=7

at least nine people were killed and up to 37 wounded

i343 TRIGGER: wounded
AGENT: the FAZ (Zaire army): ID=53: CHAIN=53: ORGANISATION
VICTIM: up to 37: ID=66: CHAIN=66: PERSON
CAUSE: –
PLACE: Bukavu: ID=70: CHAIN=17: LOCATION
TIME: Monday: ID=69: CHAIN=7

there are nine confirmed dead and 37 wounded

i346 TRIGGER: wounded
AGENT: –
VICTIM: 37 wounded: ID=86: CHAIN=78: PERSON
CAUSE: –
PLACE: –
TIME: –

Table 3. Indefinitely quantified NPs as VICTIM arguments

This is one case where indirect anaphoric relations should not be taken as acceptable to indicate coreferential events. There were 16 (10 with figures confirmed in later arguments and including DEFEND) such event chains in the corpus, which given the discussion above, is not the right annotation. The guidelines do specify that annotators should distinguish ATTACK events from DIE and INJURE events in the same sentence; however, judging by the results of the analysis, this should have been made even clearer and emphasised to the annotators in training. The difficulty arises because INJURE and DIE events are often consequences of an attack, which makes them difficult to distinguish. In addition, in ATTACK events the focus is on the attacker doing the attacking regardless of who is affected, whereas in INJURE and DIE events, the emphasis is on the victim dying or receiving an injury and the event therefore depends on the 'recipient', which must be the same person for coreferential DIE or INJURE event mentions. As the news texts in the corpus report a set of related events per text, when the annotator marked coreference between these event mentions, they were indicating, for example, that the injuring of up to 37 people caused by the Zaire Army planes’ bombs in Bukavu and the injuring of 37 people caused by the Zaire Army planes’ bombs are related to the same INJURE event. They are indicating that these are not related to another INJURE event in the same text, such as the injuring of at least 6 people caused by the Zaire Army planes’ bombs in Shubundi. Given the nature of the coreferential relation, this is not strictly what they should have annotated. An alternative
interpretation is to consider these event mentions as part of the same chain because they would be coreferential if they were annotated as a different type of event (e.g. ATTACK).

4.2 Slot-related issues

A large number of events do not have all their argument slots filled. This is due to a variety of reasons, such as the annotation strategy adopted, the class of trigger and style/readability issues in the news texts. The NP4E annotation strategy instructed annotators to fill slots with arguments from the same sentence, unless information in surrounding sentences helped to disambiguate the event from a similar one. Therefore, if the only arguments available in the same sentence as one particular CONTACT trigger are, for example, AGENT and TOPIC, but TIME is available in the next sentence which also contains a trigger for this event, the first trigger should be assigned the AGENT and TOPIC arguments, and the second trigger should take the TIME. Once marked as coreferential, the slots for these triggers will be linked together. However, the annotators used information from surrounding sentences fairly often, although not particularly systematically (see Section 5).

Certain slots are more likely to be filled for certain types of trigger. For example, a verb trigger usually has at least a subject from its sentence, which carries out the action or can be a target of another action. Noun triggers can legitimately have fewer slots filled because the information that would be arguments for a verb trigger are incorporated in the NP itself (Thursday’s meeting, Peru standoff). However, annotators sometimes chose to fill as many slots as possible by using information from the trigger itself, so the TIME and PLACE attributes are filled by Thursday and Peru, respectively. Related to this, the arguments assigned to TIME and PLACE slots often feature as adjuncts in their sentences, and therefore are not obligatory elements of sentences, unlike, for example, subjects and objects (where taken). This helps to account for cases where TIME and PLACE slots are not filled and the fact that participants such as AGENT, ATTACKER and TARGET are more likely to be present.

The concept of unfilled arguments can be linked to the idea to implicit arguments (or open roles) in discourse [18], which act as a cohesive device and are considered a type of anaphor referring to another mention in the text. This is closely related to the arguments certain syntactic structures demand and those which are optional or not present. Therefore it is of interest as most of our triggers representing events are verbs and nouns (including nominalisations). Whittemore et al.’s work [18] falls within the framework of DRT (Discourse Representation Theory [19]), an exploration of which is beyond the scope of the current paper. However, in future it would be interesting to analyse examples from the NP4E corpus using this version of DRT to establish the extent to which this can account for the lack of filled argument slots in the event chains.

There are also style and readability issues relating to the repetition of information and not all sentences containing an event mention contain all its related information because the text becomes repetitive and not particularly easy to read. Slots are more likely to be repeated when arguments help to elaborate or disambiguate the event, or to remind the reader of things mentioned previously in the text. Repetition is also likely to occur in direct speech describing the event, because the speech is produced separately and the speaker is not aware of how it will be incorporated into the text. For the empty
slots, it is necessary to propagate the missing information from other mentions of that event, where these are present. An interesting observation is that the first mention of the event does not necessarily contain all its arguments, which may be introduced later in the text. This is particularly true for events first mentioned in headlines, common in the NP4E corpus due to its composition.

Our analysis showed that two event chains (1.92%) had none of the argument slots filled for their triggers at all. A further 7 (6.73%) chains had only one argument present for any slots filled and therefore a relation could not be established. The original text was checked to establish whether this was a decision made by the annotator or whether none of the arguments were actually present in the same sentence as the trigger. The triggers for all event mentions within these chains, with the exception of two (one complete event chain), were all nouns. As mentioned above, noun triggers are theoretically less likely to have argument slots filled because they are not obligatory in the original sentence structure. The texts revealed that events can be talked about naturally without mentioning the arguments that the NP4E project annotated. In these cases, marking event coreference consisted of marking coreferential relationships between NPs representing the same event mentions. The following noun triggers (in bold) were annotated as coreferential even though no argument slots in the chain were filled. There are other clues in the sentences that the triggers refer to the same talks, related to their progression, although these did not fit into any of the argument slots defined for CONTACT events:

  Peru’s Fujimori says hostage talks still young.
  ...the President said talks to free them were still in their preliminary phase.
  "We cannot predict how many more weeks these discussions will take.”
  "We are still at a preliminary stage in the conversations.”
  Fujimori said he hoped Nestor Cerpa would personally take part in the talks when they resume on Monday at 11am.

In the last example above, the annotator was correct in not assigning Nestor Cerpa in the AGENT slot as the text does not specify at this point that he was definitely involved in the talks. Monday at 11am was not assigned to the TIME slot because the event marked here is ongoing talks, rather than a specific instance of one meeting within the talks (also annotated in this text) and so this time is not accurate for this event chain.

Another slot-related issue identified is related to CONTACT events, which have a participant slot AGENT indicating the parties involved. As these events can be extensive, involving 2 or more sides, representatives, independent mediators, leaders, etc., not all participants are described in every mention and each trigger engages only some of the participants. Therefore, it is necessary to exploit the coreferential links between event mentions to establish all the parties involved. This can be seen in Table 2, where the NP rebels as an AGENT only appears once therefore is not coreferential with any of the other fillers, and the relation between them is difficult to establish by looking at that event chain alone. This leads us on to the point that when annotating clusters of news texts, different texts may present the same larger CONTACT event from other perspectives and mention different participants. Therefore it may be necessary to exploit cross-document coreference (i.e., to consider the same events mentioned in related documents) to ensure that all parties involved in an event are identified. Although cross-document coreference
Table 4. Agreement on annotation of event chains

<table>
<thead>
<tr>
<th></th>
<th>Bukavu Annotated</th>
<th>Matched</th>
<th>Unmatched</th>
<th>Peru Annotated</th>
<th>Matched</th>
<th>Unmatched</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ann. 1</td>
<td>13</td>
<td>12</td>
<td>1</td>
<td>14</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>Ann. 2</td>
<td>14</td>
<td>12</td>
<td>2</td>
<td>12</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>Total</td>
<td>27</td>
<td>24</td>
<td>13</td>
<td>26</td>
<td>8</td>
<td>18</td>
</tr>
</tbody>
</table>

is beyond the scope of this paper (and the NP4E corpus in its present state), we should bear this potential necessity in mind for future work in extending the corpus.

5 Comparison of annotations

In addition to issues discussed in the previous section, the analysis identified instances of disagreement between annotations on the same text. This section compares the chains marked by the two annotators, focusing on triggers and argument slots. Some general annotation errors, such as an element being missed from a chain or two obviously non-coreferential mentions being annotated were also observed. These are not systematic errors and therefore we do not discuss them further. Two texts from the Bukavu bombing cluster (1,300 words) and two from the the Peru hostages cluster (1,540 words) were annotated by both annotators. Table 4 shows the agreement between annotators on marking coreferential event chains in the two clusters. Annotator 1 marked 27 chains in total and annotator 2 marked 26.

The number of chains marked by the two annotators in each Bukavu text was similar. The chains in each Peru text proved much more difficult to agree on because of the extensive amount of CONTACT event mentions involving different people and the fact that annotators sometimes focused on different types of CONTACT events. There is more agreement in general on ATTACK events than on CONTACT events; ATTACK events were most common in the Bukavu texts and CONTACT were most common in the Peru texts. If we consider the two clusters together, we have agreement in 60.34% of cases. Each cluster needs to be explained separately because they had very different results. For the Bukavu texts, annotator 1 did not mark 2 CONTACT chains marked by annotator 2, but these chains did incorporate a mention which annotator 1 considered unique. Annotator 1 also marked an ATTACK chain not identified by annotator 2. For this cluster, 88.89% of chains matched.

In comparison, the agreement on annotation of event chains in the Peru texts was only 30.77% agreement. As mentioned above, there were differences in the number of chains marked by each annotator in each text. In Peru 1, annotator 1 marked 10 chains and annotator 2 marked 6; in Peru 2 these figures were 4 and 6, respectively. Part of the reason for this high level of disagreement is that not all mentions of the hostage crisis were considered by both annotators to be markable. In addition, one annotator marked general CONTACT events such as reporting and telling, whereas the other only annotated those CONTACT events which were a more central part of the security situation, such as meetings between the different participants.
He said three jets dropped several bombs in Bukavu, the capital of South Kivu province.

<table>
<thead>
<tr>
<th>TRIGGER:</th>
<th>Annotator 1</th>
<th>Annotator 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATTACKER:</td>
<td>dropped</td>
<td>several bombs</td>
</tr>
<tr>
<td>MEANS:</td>
<td>three jets, several bombs</td>
<td>three jets</td>
</tr>
<tr>
<td>PLACE:</td>
<td>Bukavu, the capital of South Kivu province</td>
<td>Bukavu, the capital of South Kivu province</td>
</tr>
<tr>
<td>TARGET:</td>
<td>Bukavu, the capital of South Kivu province</td>
<td>Bukavu, the capital of South Kivu province</td>
</tr>
<tr>
<td>TIME:</td>
<td>14.00 GMT</td>
<td>–</td>
</tr>
</tbody>
</table>

Table 5. Different triggers marked for the same event mention

A further explanation is that, as in the Bukavu texts, where one annotator marked fewer chains, the single mentions they marked could be found in the other annotator’s chains. This happens when there are very similar events or related events which can also be interpreted as one larger event. For example, in the sentence below, one annotator considered the triggers tied up and tossed as two separate ATTACK events because of the different triggers and the other considered them as one because the slots had exactly the same arguments:

On Saturday, armed men calling themselves MRTA rebels burst into a Lima textile factory, tied up the employees and then tossed grenades into the plant, sparking a huge fire.

This type of interpretation accounts for 8 of the 17 unmatchig chains and indicates that stricter guidance must be given to annotators about what should be considered separate events, what counts as one event, and at what level of granularity the annotation should be performed. Perhaps it would also be beneficial to annotate event structures where all related events and sub-events contributing to one larger event are considered. If we can see how similar events relate to one another, we may be able to better explain why the two were considered as both single and coreferential mentions by different annotators. However, we would still need to define how to deal with mentions of events which are very similar for their inclusion in this overall structure.

Related to this, different triggers were selected as best representing the event under discussion by different annotators for the same chain in several cases. The guidelines point out that there can be more than one potential trigger for an event, and annotators were instructed to choose just one, which sometimes led to disagreement. Since the NP4E annotation was completed, other researchers have recognised that anaphoric expressions can refer to multiple antecedents and have incorporated this into their annotation [10]. The analysis of the NP4E corpus supports this finding and therefore it would be useful to acknowledge in future annotations if we are to accurately capture the range of ways events are expressed in text. Comparing the slots filled by both annotators for these different triggers, and the other triggers and arguments in the chain, we can still establish that the trigger refers to the same event (see Table 5).
Table 5 also highlights the issue that there are sometimes differences in the arguments assigned to the same slots by the two annotators. The event shown in this table is the largest and most obvious example of this, as both annotators’ chains contain 10 mentions; one chain has 7 ATTACKER fillers, and the other has 0. One annotator considered planes (planes, Zairean jets, three Zairean jet fighters, etc.) as both ATTACKER and MEANS, and the other annotator only considered them as MEANS. There were similar examples involving DEFENDER and ENTITY arguments in DEFEND events and TARGET and PLACE in ATTACK events. The guidelines specify which types of entity can be assigned to which slot, but OTHER is also offered in case the annotator cannot decide. The analysis of the annotation indicates that this use of OTHER should be clarified. If there are genuine cases where the same argument can be used for more than one different slot, this should also be made clearer in the guidelines. However, the guidelines were produced in various iterations, with detailed discussions between annotators and revisions were made accordingly in an attempt to achieve the highest possible agreement.

In many of the chains, the annotators filled different numbers of slots for each trigger. On checking the texts, we saw that where one annotator did not assign an argument this is because the argument was not present in the trigger’s sentence. The other annotator used the surrounding sentences to fill these slots so that the maximum possible information about the event in all mentions was included. Unfortunately, the annotators did not do this consistently; it seemed to depend on the individual text and the events described within it. This was done mainly for events mentioned in the body of the news text and not the headline. Where the headline contained the first mention of the event, the annotator tended to assign arguments from the headline only. This may be due to the organisation of news texts, where the headline is set apart and can be considered a summary of the rest of the text.

6 Conclusions and future work

This paper presented an investigation of coreferential event chains starting from the relations between arguments filling the same slot. Our hypothesis was that the arguments in coreferential event mentions would also be coreferential. However, the analysis of the NP4E corpus showed that this is not a necessary feature. In 70% of the event chains annotated, indirect anaphoric relations were present between the arguments. This finding has important implications for researchers developing automatic methods for event coreference resolution because it shows that indirect anaphora needs to be tackled before we can deal with event coreference successfully. We also observed that many argument slots were not filled, but that annotators could still identify coreference between event mentions.

We found disagreement between the annotators on which triggers were marked and which slots were filled, despite the fact that the guidelines were developed through extensive discussions and the annotation happened in several iterations. This highlights the difficulty of the annotation task and the need for more explicit guidelines and a more precise definition of event. One way of helping to refine this definition and to distinguish events from similar ones is to use an approach inspired by Bejan and Harabagiu [17]. As
well as annotating coreferential event mentions, event structures including all sub- and related events and the relations between them would be annotated. Despite increasing the complexity of the annotation process, this should give us more comprehensive information about events.

Related to this, although cross-document coreference is beyond the scope of this paper (and the NP4E corpus in its present state) it is an issue to bear in mind for future work in extending the corpus. A layer of cross-document coreference annotation for both arguments and events could be added to make sure that all arguments of event mentions (particularly participants) are captured.

In answer to the question posed by the title of our paper, we can conclude that coreferential arguments do make event mentions coreferential, but so do arguments in indirect anaphoric relationships in many more cases.

References

2. NIST: ACE. http://www.itl.nist.gov/iad/igi/tests/ace/


