Multi-Topic Multi-Document Summarization

UTIYAMA Masao
Communications Research Laboratory
588-2, Iwaoka, Nishi-ku, Kobe,
Hyogo 651-2492, Japan
mutiyama@crl.go.jp

HASIDA Kôiti
Electrotechnical Laboratory
1-1-4, Umezono, Tukuba,
Ibaraki 305-8568, Japan
hasida@etl.go.jp

Abstract

Summarization of multiple documents featuring multiple topics is discussed. The example treated here consists of fifty articles about the Peru hostage incident for December 1996 through April 1997. They include a lot of topics such as opening, negotiation, ending, and so on. The method proposed in this paper is based on spreading activation over documents syntactically and semantically annotated with GDA (Global Document Annotation) tags. The method extracts important documents and important parts therein, and creates a network consisting of important entities and relations among them. It also identifies cross-document coreferences to replace expressions with more concrete ones. The method is essentially multilingual due to the language-independence of the GDA tagset. This tagset can provide a standard format for the study on the transformation and/or generation stage of summarization process, among other natural language processing tasks.

1 Introduction

A large event consists of a number of smaller events. These component events are usually related but such relations may not be strong enough to define larger topics. For example, a war may consist of opening, battles, negotiations, and so on. These relatively independent events are considered to be topics by themselves and would accordingly be reported in multiple news articles.

Summarization of such a large event, or multiple documents about multiple topics, is the concern of this paper. Summarization of multiple documents containing multiple topics is an unexplored research issue. Some previous studies on summarization (McKeown and Radev, 1995; Barzilay et al., 1999; Mani and Bloedorn, 1999) deal with multiple documents about a single topic, but not about multiple topics[1].

In order to summarize multiple documents with multiple topics, one needs a general, semantics-oriented method for evaluating importance. Summarization of a single document may largely exploit the document structure. As an extreme example, the first paragraph of a newspaper article often serves as a summary of the entire article. On the other hand, summarization of multiple documents in general must be more based on their semantic structures, because there is no overall consistent document structure across them.

Selection of multiple important topics (not keywords) for multiple-topic summarization has not yet been really addressed in the previous literature. The present paper proposes a method, based on spreading activation, for extracting important topics and important documents. Another method proposed which is useful for grasping the overview of multiple documents is visualization of important entities mentioned and relationships among them. Visualization of relationships among keywords has been studied in the context of information retrieval (Niwa et al., 1997; Sanderson and Croft, 1999), but to the authors' knowledge the present study is the first to address such visualization in the context of summarization. Of course a concise summary of the entire set of multiple documents can be obtained by recovering sentences from important entities and their relationships as demonstrated in section 3.3.

The present study assumes documents annotated with GDA (Global Document Annota-

[1] Maybury (1999) discusses summarization of multiple topics, but in his study the summaries are made from an event database but not from documents.

tion) tags (Hasida, 1997; Nagao and Hasida, 1998). Since the GDA tagset is designed to be independent of any particular natural language, the proposed method is essentially multilingual. Another merit of using annotated documents is that we can separate the analysis phase from the whole process of summarization so that we can focus on the latter, generation phase of summarization process. Annotated documents can also be useful for a common input format for the study of summarization, among other natural language processing tasks.

2 The GDA Tagset

GDA is a project to make on-line documents machine-understandable on the basis of a linguistic tagset, while developing and spreading technologies of content-based presentation, retrieval, question-answering, summarization, translation, among others, with much higher quality than before. GDA thus proposes an integrated global platform for electronic content authoring, presentation, and reuse. The GDA tagset[2] is an XML (eXtensible Markup Language) instance which allows machines to automatically infer the semantic and pragmatic structures underlying the raw documents.

Under the current state of the art, GDA-tagging is semiautomatic and calls for manual correction by human annotators; otherwise annotation would make no sense. The cost involved here pays, because annotated documents are generic information contents from which to render diverse types of presentations, potentially involving summarization, narration, visualization, translation, information retrieval, information extraction, and so forth. The present paper concerns summarization only, but the merit of GDA-tagging is not at all restricted to summarization, and that is why it is considered reasonable to assume GDA-tagged input here.

2.1 Syntactic structure

An example of a GDA-tagged sentence is shown in Figure 1. `<su>` means sentential unit. `<np>`, `<v>`, and `<adp>` stand for noun phrase, verb, and adnominal or adverbial phrase.

`<su>` and the tags whose name end with 'p' (such as `<adp>` and `<vp>`) are called *phrasal tags*. In a sentence, an element (a text span from a begin tag to the corresponding end tag) is usually a syntactic constituent. The elements enclosed in phrasal tags are called *phrasal elements*, which cannot be the head of larger elements. So in Figure 1 'flies' is specified to be the head of the `<su>` element and 'like' the head of the `<adp>` element.

[2] http://www.etl.go.jp/etl/nl/GDA/tagset.html

```
<su>
  <np>Time</np>
  <v>flies</v>
  <adp>
    like
    <np>an arrow</np>
  </adp>
  .
</su>
```

Figure 1: A GDA-tagged sentence.

2.2 Coreferences and Anaphora

Each element may have an identifier as the value for the `id` attribute. Coreferences, including identity anaphora, are annotated by the `eq` attribute, as follows:

```
<np id="j0">John</np> beats
<adp eq="j0">his</adp> dog.
```

When the shared semantic content is not the referent but the type (kind, set, etc.) of the referents, the `eq.ab` attribute is used like the following:

```
You bought a <np id="c1">car</np>.
I bought <np eq.ab="c1">one</np>,
too.
```

A zero anaphora is encoded as follows:

```
Tom visited <np id="m1">Mary</np>.
He had <v iob="m1">brought</v> a
present.
```

`iob="m1"` means that the indirect object of *brought* is element whose `id` value is `m1`, that is, *Mary*.

Other relations, such as `sub` and `sup`, can also be encoded. `sub` represents subset, part, or element. An example follows:

```
She has <np id="b1">many
books</np>.
<namep sub="b1">‘‘Alice’s
```

```
Adventures in Wonderland''</namep>
is her favorite.
```

`sup` is the inverse of `sub`, i.e., includer of any sort, which is superset as to subset, whole as to part, or set as to element.

Syntactic structures and coreferences are essential for the summarization method described in section 3. Further details such as semantics, coordination, scoping, illocutionary act, and so on, are omitted here.

3 Multi-Document Summarization

3.1 Spreading activation

A set of GDA-tagged documents is regarded as a network in which nodes roughly correspond to GDA elements and links represent the syntactic and semantic relations among them. This network is the tree of GDA elements plus cross-reference (via `eq`, `eq.ab`, `sub`, `sup`, and so on) links among them. Cross-reference links may encompass different documents. Figure 2 shows a schematic, graphical representation of the network.

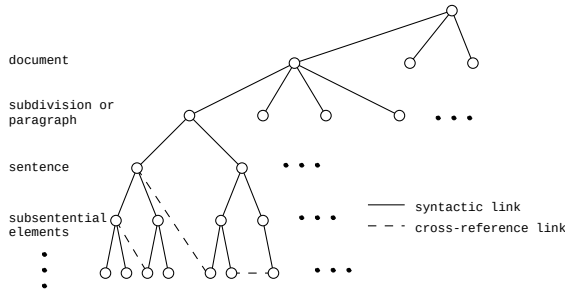

Figure 2: Multi-document network.

Spreading activation is carried out in this network to assess the importance of the elements. Spreading activation has been applied to summarization of single GDA-tagged documents (Hasida et al., 1987; Nagao and Hasida, 1998). The main conjecture of the present study is that the merit of spreading activation in that it evaluates importances of semantic entities is greater in summarization of multiple documents with multiple topics, because summarization techniques using document structures do not apply here, as mentioned earlier.

To fit the semantic interpretation, activations spread under the condition that coreferent elements should have the same activation value. The algorithm[3] is shown in Figure 3. Here the external input $c(i)$ to node i represents *a priori* importance of i, which is set on an empirical basis; for instance, an entity [4] referred to in the title of an article tend to be important, and thus $c(i)$ should be relatively large for the corresponding node i. The weight $w(i,j)$ of another kind of link from node i to node j may also be set empirically, but it is fixed to a uniform value in the present work. Let $E(i)$ be the equivalence class of node i, that is the set of nodes which are coreferent with i (linked with i via `eq` relationships). Condition

$$\sum_{k \in E(i)} \sum_{j \notin E(i)} w(k,j) \leq 1$$

should be satisfied in order for the spreading activation to converge. This condition is satisfied if we treat each equivalence class of nodes as a virtual node while setting the weights of other types of links to be $\frac{1}{D}$, where D is the maximum degree of equivalence classes:

$$D = \max_i \sum_{k \in E(i)} \sum_{j \notin E(i)} \delta_{kj}$$

where δ_{kj} is 1 if there is a link between node k and node j, otherwise it is 0.

The score $score(i)$ of node i is calculated by summing the activation values of all the nodes under node i in the syntactic tree structure:

$$score(i) = a(i) + \sum_{j \in ch(i)} score(j) \qquad (1)$$

where $a(i)$ is the activation value of node i and $ch(i)$ is the set of child nodes of node i. $ch(i)$ is empty if node i is a leaf node, or a word. This score is regarded as the importance of node i.

3.2 Extraction of important documents and sentences

Extraction of important documents is simple once the scores of the nodes in the network are obtained. Sorting the document nodes according to their scores and extracting higher-ranked ones is sufficient for the purpose.

[3] Another spreading activation algorithm is discussed by Mani and Bloedorn (1999). The comparison is a future work.

[4] We use the terms 'entity', 'node', and 'element' interchangeably.

Variables:

`N`: number of nodes.

`D`: maximum out-degree of equivalence classes.

`c(i)`: external input to node `i`.

`w(i,j)`: weight of the link from node `i` to node `j`:

0	if not connected,
1	if connected via `eq`,
1/D	otherwise.

`a(i)`: activation value of node `i`. The initial value is 0. `a(i)` is the sum of all *a(j,i)*.

`a(i,j)`: activation value of the link from node `i` to node `j`. The initial value is 0.

Algorithm:

```
repeat {
  for(i=0; i<N; i++){
    av = c(i);
    for(j=0; j<N; j++){
      a(j,i) = w(j,i)*(a(j) - a(i,j))
      av += a(j,i)
    }
    a(i) = av;
  }
} until convergence.
```

Figure 3: Spreading activation algorithm.

Similar procedure is used to extract important sentences from an important document. Extracted sentences are pruned according to their syntactic structures. Anaphoric expressions such as *he* or *she* are substituted by their antecedents if necessary.

An experiment has been conducted to test the effectiveness of the proposed algorithm. The example set contains fifty Japanese articles about the Peru hostage incident which continued over four months from December 1996 to April 1997. They include a lot of topics such as opening, negotiation, settlement, and so on. The GDA-tagging of these articles has involved automatic morphological analysis by JUMAN (Kurohashi and Nagao, 1998), automatic syntactic analysis by KNP (Kurohashi, 1998), and manual annotation encompassing morphology, syntax, coreference, and anaphora. The types of anaphora identified here are mainly plain coreference and zero anaphora. Cross-document coreferences among entities have been automatically identified by exact string matching. [5] They contained errors but those errors were not corrected for the experiment. Cross-document coreferences found were 'Peru'(49), 'Japan'(39), 'Peru President' (15), 'members of Tupac Amaru'(9), ... and so on, where the numbers indicate the numbers of documents which contain these expressions.

The external inputs to nodes have been defined according to the corresponding nodes: $c(i) = 10$ if node i's antecedent dominates sentences (e.g., a node coreferring with a paragraph). This sets a preference for nodes which summarize preceding sentences. $c(i) = 5$ if node i is in the title of an article, because a title is usually important. Otherwise $c(i) = 1$. These crude parameter values have been set by the authors on the basis of the investigation of summarizations of various documents.

Two important topics, the opening (first attack by Tupac Amaru) and the settlement (attack by the Peruvian government commandos), have been extracted from the four highest ranked articles, even though temporal information has not been incorporated in the algorithm. The opening article is the first article of the sample document set. However, the settlement article is the sixth last one. So mere extraction of the last article would miss the settlement.

The 25% summaries of the two articles made by extracting and pruning sentences are shown below together with their English translations:

> 日本大使公邸に武装ゲリラ、パーティーに乱入
> 銃撃、200 人が人質ーーペルー
>
> 日本、ペルーの両国関係者多数が人質にとられた。武装グループは約 20 人で、うち複数が公邸に押し入った。現在、散発的に銃撃戦が展開されているという。
>
> Armed guerrillas broke into a party at Japanese ambassador's residence. Gunshots. 200 held in hostage. — Peru.
>
> Many people from Japanese and Peruvian sides were held in hostage. The armed group consists of about twenty people, several of which broke into the ambassador's residence. It is reported that there are intermittent shootings now.

and

[5] We are planning to incorporate recent results (Bagga and Baldwin, 1998) to identify cross-document coreferences.

ペルー日本大使公邸占拠事件　人質全員解放
権力基盤回復狙いーーフジモリ大統領

フジモリ大統領は、強気の政治家であることを、日本大使公邸占拠事件の解決でみせつけた。フジモリ大統領は公邸敷地に入った。公邸訪問は、作戦の陣頭指揮を執っていることを印象付けた。なぜ、フジモリ大統領は武力行使を選択したのか。政治危機の根源にある公邸事件を、武力で解決することで、政治主導権の回復を狙ったといえる。

Japanese ambassador's residence possession incident in Peru. All hostages released. Aim at recovering his power basis — President Fujimori.

President Fujimori demonstrated himself as a strong politician by resolving the Japanese ambassador's residence possession incident. He entered the residence site. This visit to the residence impressed that he was leading the operation himself. Why did he choose to resort to arms? We can say that he aimed at recovering his political leadership by resolving through military power the residence incident, which is at the root of the political crisis.

3.3 Entity-relation graph

The score $score(i, j)$ of a relation between two entities i and j, is defined by:

$$score(i,j) = |E(i)|a(i) + |E(j)|a(j) + \sum_{s \in S(E(i)) \cap S(E(j))} score(s) \quad (2)$$

where $S(E(i))$ is the set of sentence nodes which dominate one of the nodes in $E(i)$ and $|E(i)|$ is the number of nodes in $E(i)$. $E(i)$, $a(i)$, and $score(s)$ have been defined in Section 3.1. $|E(i)|a(i)$ is an analogy of '$tf \times idf$', which is a measure of term importance widely used in information retrieval.

If $score(i, j)$ is sufficiently large, then $S(E(i)) \cap S(E(j))$ (the sentences containing both the entities) can constitute a cross-document summary concerning i and j[6].

An entity-relation graph (E-R graph) is made of the relations highly ranked in terms of the score defined in (2). Figure 4 shows the E-R graph made of the top eleven relations extracted from the articles about Peru hostage incident. The numbers near the lines represent the ranks of the relations.

[6]Coreference chains are used to summarize single documents by Azzam et al. (1999).

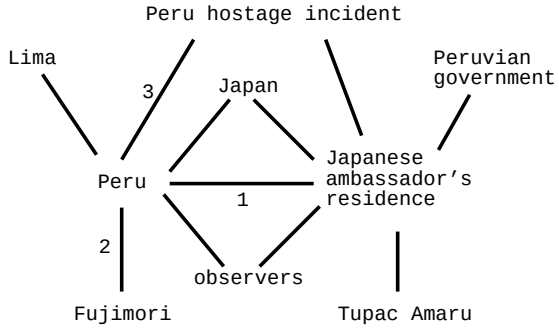

Figure 4: E-R graph of Peru hostage incident.

The top-ranked relation was the one between *Peru* and *Japanese ambassador's residence*. Three sentences extracted from the eight sentences which contained both of the entities were as follows.[7] They were listed in chronological order which was identified by the date information in the articles.

1. ペルーからの報道によると、首都リマ市にある日本大使公邸が 17 日、左翼ゲリラとみられる武装グループに襲撃され、日本、ペルーの両国関係者多数が人質にとられた。

   According to reports from Peru, **on the 17th** the Japanese ambassador's residence in Lima, the capital, was attacked by **an armed groups, allegedly leftist guerrillas**, and many people from both Japanese and Peruvian sides were held in hostage.

2. ペルーの日本大使公邸で発生した武装ゲリラによる人質事件で政府は 18 日、ペルー政府に対し人質の安全確保を要請するとともに、外務省の堀村隆彦中南米局審議官を同夜、現地に派遣した。

   Concerning **the hostage incident at the Japanese ambassador's residence caused by armed guerrilla**, **on the 18th** the government requested the Peruvian government to assure the safety of the hostages, and sent Mr. HORIUTI Takahiko, coordinator, Division of Middle and South America. ···

3. 22 日、ペルーの日本大使公邸への突入作戦を成功させたことで、フジモリ大統領の政治的威信は再び回復に向かうことになるだろう。

   President Fujimori's political authority will recover because he succeeded in the operation to break into the Japanese ambassador's residence in Peru **on the 22nd**.

These sentences, extracted from different articles, have been paraphrased on the basis of

[7]These sentences were selected manually to demonstrate the possibility of cross document summarization based on coreference.

coreferences. Since the name of the guerilla group is not identified in the beginning of the incident, the expression '左翼ゲリラとみられる武装グループ' (armed group which seems to be leftist guerrillas) is used in the first sentence there. This expression has been replaced with '左翼ゲリラ (トゥパク・アマル)' (leftist guerrillas (Tupac Amaru)) by using cross-document coreferences. The equivalence of the first sentence and the first noun phrase of the second sentence, 'ペルーの日本大使公邸で発生した武装ゲリラによる人質事件' (the hostage incident caused by armed guerrillas at the Japanese ambassador's residence in Peru), were properly detected and was replaced by another expression because the equivalence of events across possibly different documents (McKeown et al., 1999; Barzilay et al., 1999) has been also detected by comparing predicate-argument structures of relevant sentences. Date expressions such as '17 日' (the 17th) have been augmented like '1996 年 12 月 17 日' (Dec. 17, 1996). The resulting passages are below (underlines indicating paraphrases), together with their English translations (boldface indicating paraphrases):

1. ペルーからの報道によると、首都リマ市にある日本大使公邸が 1996 年 12 月 17 日、左翼ゲリラ (トゥパク・アマル) に襲撃され、日本、ペルーの両国関係者多数が人質にとられた。

   According to news from Peru, **on December 17, 1996** the Japanese ambassador's residence in Lima, the capital, was attacked by **leftist guerrillas (Tupac Amaru)**, and many people from both Japanese and Peruvian sides were held in hostage.

2. その人質事件で政府は 12 月 18 日、ペルー政府に対し人質の安全確保を要請するとともに、外務省の堀村隆彦中南米局審議官を同夜、現地に派遣した。

   Concerning **the hostage incident**, the government requested the Peruvian government to assure the hostages' safety **on December the 18th**, and sent Mr. HORIUTI Takahiko, coordinator, Division of Middle and South America, Ministry of International Affairs, to Peru on that night.

   ...

3. 1997 年 4 月 22 日、ペルーの日本大使公邸への突入作戦を成功させたことで、フジモリ大統領の政治的威信は再び回復に向かうことになるだろう。

   President Fujimori's political authority will recover because he succeeded in the operation to break into the Japanese ambassador's residence in Peru **on April 22, 1997**.

4 Discussion

4.1 Evaluation

Evaluation of multi-document summarization calls for far greater cost than that of single-document summarization. Testbeds for evaluation of multi-document summarization have not been developed yet. So the present evaluation is limited to the sample set of articles mentioned above, but the obtained results suggest general applicability of the proposed method and supports the conjecture that spreading activation is effective for multi-document multi-topic summarization.

As discussed in the previous section, the proposed method can extract important articles, that is, the opening and settlement articles, from fifty articles about Peru hostage incident. Also, an E-R graph consisting of important relations among important entities, *Peru, Japanese ambassadors' residence, Tupac Amaru*, and so on, has been successfully constructed on this basis. The above-mentioned method also uses cross-document coreferences for replacing expressions with more concrete ones.

All these are archived essentially by using information in the GDA-tagging only, but not domain-dependent knowledge such as embedded in templates for information extraction. The proposed method is hence expected to detect important documents and sentences and create an appropriate E-R graph when applied to another set of documents about multiple topics.

4.2 Transformation

The process of summarization can be decomposed into three stages (Sparck Jones, 1999):

1. source text *interpretation* to source text representation,
2. source representation *transformation* to summary text representation, and
3. summary text *generation* from summary representation.

GDA-tagged documents are regarded as source text representations. The method described above focuses on the transformation stage. Its multi-linguality comes from the multi-linguality of the stage.

5 Conclusion

Summarization of multiple documents about multiple topics has been discussed in this paper. The method proposed here uses spreading activation over documents syntactically and semantically annotated with GDA tags. It is capable of:

- extraction of the opening and settlement articles from fifty articles about a hostage incident,
- creation of an entity-relation graph of important relations among important entities,
- extraction and pruning of important sentences, and
- substitution of expressions with more concrete ones using cross-document coreferences.

The method is essentially multilingual because it is based on GDA tags and the GDA tagset is designed to address multilingual coverage. Since this tagset can embed various linguistic information into documents, it could be a standard format for the study of the transformation and/or generation stage of document summarization, among other natural language processing tasks.

References

Saliha Azzam, Kevin Humphreys, and Robert Gaizauskas. 1999. Using coreference chains for text summarization. In *ACL'99 Workshop on Coreference and Its Applications*, pages 77–84.

Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *COLING-ACL'98*, pages 79–85.

Regina Barzilay, Kathleen R. McKeown, and Michael Elhadad. 1999. Information fusion in the context of multi-document summarization. In *ACL'99*, pages 550–557.

Kôiti Hasida, Syun Ishizaki, and Hitoshi Isahara. 1987. A connectionist approach to the generation of abstracts. In Gerard Kempen, editor, *Natural Langauge Generation: New Results in Artificial Intelligence, Psychology, and Linguistics*, pages 149–156. Martinus Nijhoff.

Kôiti Hasida. 1997. Global Document Annotation. In *NLPRS'97*, pages 505–508.

Sadao Kurohashi and Makoto Nagao. 1998. Japanese morphological analysis system JUMAN manual.

Sadao Kurohashi. 1998. Japanese syntactic analysis system KNP manual.

Inderjeet Mani and Eric Bloedorn. 1999. Summarizing similarities and differences among related documents. In Inderjeet Mani and Mark T. Maybury, editors, *ADVANCES IN AUTOMATIC TEXT SUMMARIZATION*, chapter 23, pages 357–379. The MIT Press.

Mark T. Maybury. 1999. Generating summaries from event data. In Inderjeet Mani and Mark T. Maybury, editors, *ADVANCES IN AUTOMATIC TEXT SUMMARIZATION*, chapter 17, pages 265–281. The MIT Press.

Kathleen McKeown and Dragomir R. Radev. 1995. Generating summaries of multiple news articles. In *SIGIR'95*, pages 74–82.

Kathleen R. McKeown, Judith L. Klavans, Vasileios Hatzivassiloglou, Regina Barzilay, and Elezar Eskin. 1999. Towards multidocument summarization by reformulation: Progress and prospects. In *AAAI-99*, pages 453–460.

Katashi Nagao and Kôiti Hasida. 1998. Automatic Text Summarization Based on the Global Document Annotation. In *COLING-ACL'98*, pages 917–921.

Yoshiki Niwa, Shingo Nishioka, Makoto Iwayama, Akihiko Takano, and Yosihiko Nitta. 1997. Topic graph generation for query naviagation: Use of frequency classes for topic extraction. In *NLPRS'97*, pages 95–100.

Mark Sanderson and Bruce Croft. 1999. Deriving concept hierarchies from text. In *SIGIR'99*, pages 206–213.

Karen Sparck Jones. 1999. Automatic summarizing: factors and directions. In Inderjeet Mani and Mark T. Maybury, editors, *ADVANCES IN AUTOMATIC TEXT SUMMARIZATION*, chapter 1, pages 1–12. The MIT Press.