

# An Anaphora Resolution-Based Anonymization Module

M. Poesio,\* M. A. Kabadjov,\* P. Goux,\* U. Kruschwitz\*, E. Bishop† and L. Corti†

University of Essex

\*Department of Computer Science / Language and Computation Group

†UKDA

Colchester C04 3SQ, United Kingdom

## Abstract

Growing privacy and security concerns mean there is an increasing need for data to be anonymized before being publically released. We present a module for anonymizing references implemented as part of the SQUAD tools for specifying and testing non-proprietary means of storing and marking-up data using universal (XML) standards and technologies. The tool is implemented on top of the GUITAR anaphoric resolver.

## 1. Introduction

Growing privacy and security concerns mean there is an increasing need for data to be anonymized before being publically released. Providing tools to facilitate the task is one of the goals of the Smart Qualitative Data SQUAD project, one of whose objectives is to use natural language processing technology—specifically, the LT-XML tools<sup>1</sup> developed by the University of Edinburgh’s Language and Technology Group—to develop and implement user-friendly tools for semi-automating processes to prepare qualitative data for traditional digital archiving and other types of processing. The tools developed as part of the project should make it possible for Social Sciences researchers to access data such as the transcripts of interviews stored in the University of Essex’s Data Archive. However, the names of the individuals who agreed to participate in the interview need to be anonymized, possibly in an automatic form.

In this poster we present preliminary work on an anonymization tool developed as part of the SQUAD project. Like the rest of the software developed in the project, the anonymization tool is designed to work off the LT-XML tools and to interface with the NITE XML TOOLKIT (NXT). The key idea is to take advantage of an existing anaphora resolution system also designed to interface with the LT-XML tools, the GUITAR 3.1 system (Poesio and Kabadjov, 2004; Poesio et al., 2005), which we are already using for summarization (Steinberger et al., 2005). An anonymization tool based on an anaphoric / coreference resolver could potentially simplify the task of anonymization by eliminating the need to identify all possible forms used to mention a particular individual. This experiment would also provide us with a different way of evaluating GUITAR.

In this paper, we briefly describe GUITAR, then present the anonymization algorithm, and discuss future work.

## 2. GUITAR 3.1

GUITAR is an anaphora resolution system designed to be high precision, modular, and usable as an off-the-shelf component of a NLP pipeline such as the LT-XML tools.

### 2.1. Input

GUITAR takes XML input in a format called MAS-XML,

which augments to produce output also in XML format. It can work with a variety of preprocessing tools ranging from simple POS taggers to chunkers (such as LT-CHUNK) to full parsers (an interface to Charniak’s parser has been implemented), provided that their output can be converted into MAS-XML format (typically, by heuristic methods). These features makes GUITAR very suitable for the intended application, in which it will work as a component for a preprocessing module whose output will then be manually edited for final corrections using NXT.

MAS-XML is illustrated in Figure 2, which shows the type of input GUITAR expects for a text like the one in Figure 1. At a minimum, GUITAR expects the text to have been tokenized and POS-tagged, and sentences and nominal phrases (NES) to have been identified. The system can also take advantage of other types of information if available—e.g., about grammatical function, or about named entity types.

```
My grandpa Gaunting married when my mother was -  
just under ten.  
So - he remarried.  
And my mother calls her Doris as well.
```

Figure 1: An example of raw text

### 2.2. Anaphora Resolution Algorithms

GUITAR uses an implementation of the MARS pronoun resolution algorithm (Mitkov, 1998) to resolve personal and possessive pronouns. The system resolves definite descriptions using a partial implementation of the algorithm proposed in (Poesio and Vieira, 1998), augmented with a statistical classifier to identify discourse-new definite descriptions (Poesio et al., 2005). Finally, GUITAR 3.1 also includes an implementation of the shallow algorithm for resolving coreference with proper names proposed by (Bontcheva et al., 2002).

Whenever GUITAR identifies an anaphoric relation, it adds to its output a new *aante* element specifying a possible anchor for the anaphoric expression participating in the relation; GUITAR never deletes anything from its input. For example, an ideal result for the input in Figure 2 would be for GUITAR to recognize that *my grandpa Gaunting* and

<sup>1</sup><http://www.ltg.ed.ac.uk/software/xml/>

```

<s id="s120">
  <unit id="unit198">
    <ne id="ne312">
      <W Lpos="PRP$" >My</W>
      <W Lpos="NN" >grandpa</W>
      <W Lpos="NNP" >Gaunting</W>
    </ne>
    <ve id="ve162" gId="nv162">
      <W Lpos="VBD" >married</W>
    </ve>
    <sbar id="sbar37">
      <W Lpos="WRB" >when</W>
      <unit id="unit199">
        <ne id="ne313">
          <W Lpos="PRP$" >my</W>
          <W Lpos="NN" >mother</W>
        </ne>
        <W Lpos="AUX" >was</W>
        <W Lpos=":" >-</W>
        <W Lpos="RB" >just</W>
        <W Lpos="IN" >under</W>
        <ne id="ne314">
          <W Lpos="CD" >ten</W>
        </ne>
      </unit>
    </sbar>
  </unit>
</s>
<s id="s121">
  <unit id="unit200">
    <W Lpos="RB" >So</W>
    <W Lpos=":" >-</W>
    <ne id="ne315">
      <W Lpos="PRP" >he</W>
    </ne>
    <ve id="ve163" gId="nv163">
      <W Lpos="VBD" >remarried</W>
    </ve>
  </unit>
</s>
<s id="s122">
  <unit id="unit201">
    <W Lpos="CC" >And</W>
    <ne id="ne316">
      <W Lpos="PRP$" >my</W>
      <W Lpos="NN" >mother</W>
    </ne>
    <ve id="ve164" gId="nv164">
      <W Lpos="VBZ" >calls</W>
    </ve>
    <ne id="ne317">
      <W Lpos="PRP$" >her</W>
      <W Lpos="NNP" >Doris</W>
    </ne>
    <W Lpos="RB" >as</W>
    <W Lpos="RB" >well</W>
  </unit>
</s>

```

Figure 2: The same text in MAS-XML format

*he* are mentions of the same discourse entity, as are the two mentions of *my mother*; and to add as a result to its input the anaphoric relations in Figure 3. Anaphoric relations are expressed as separate *aante* XML elements; the *current* attribute of the element specifies the index of the anaphoric expression, whereas the *antecedent* attribute of the embedded anchor element expresses the antecedent.

```

<aante current="ne315" rel="ident">
  <anchor antecedent="ne312"></anchor>
</aante>

<aante current="ne316" rel="ident">
  <anchor antecedent="ne313"> </anchor>
</aante>

```

Figure 3: Representation of Anaphoric Relations in MAS-XML

### 2.3. Performance figures

The previous version of GUITAR could resolve nominal references other than proper names (P=69.1, R=53.3) and pronouns (P= 52.9, R=52.6). In order to use the system for the anonymization application, we implemented Bontcheva et al’s proper name resolution algorithm. The newest version of the system is thus also able of resolving proper names (R= 77, P=69.6).

## 3. The Anonymization Algorithm

The preliminary version of the anonymization tool, implemented in Java, does not automatically anonymize all entities mentioned using proper names in a text. Instead, it is designed to be called by the users of the SQUAD tools whenever they decide to anonymize a particular entity, so it anonymizes one element at a time. It takes as input the output of GUITAR; the string used for the first mention of the proper name whose instances have to be anonymized; and the replacement string. It reads in GUITAR’s MAS-XML output using the Java Document Object Model (DOM) interface, and builds coreference chains for all discourse entities mentioned in the text. Then it finds the first *ne* element whose text subnode exactly matches the string specified by the user, and replaces that text subnode, and the text subnodes of all elements of that coreference chain, with the output string. It then outputs the resulting DOM tree.

For example, when called with input the output of GUITAR on the text in Figure 1, and told to replace all instances of the discourse entity first mentioned as “My grandpa Gaunting” with “XXX,” our tool will produce the output in Figure 4. This example illustrates one of the potential advantages of this tool over a simple XML-aware replacement tool: in addition to anonymizing the proper names, the tool also anonymizes the pronominal references, eliminating information about gender. (The actual usefulness of this strategy will have to be evaluated.)

## 4. Discussion

While the performance of GUITAR 3.1 is not perfect, it has been shown to be good enough to improve the performance

```

<s id="s120">
  <unit id="unit198">
    <ne id="ne312">
      <W Lpos="NNP">XXX</W>
    </ne>
    <ve id="ve162" gId="nv162">
      <W Lpos="VBD">married</W>
    </ve>
    <sbar id="sbar37">
      <W Lpos="WRB">when</W>
      <unit id="unit199">
        <ne id="ne313">
          <W Lpos="PRP$">my</W>
          <W Lpos="NN">mother</W>
        </ne>
        <W Lpos="AUX">was</W>
        <W Lpos=":">-</W>
        <W Lpos="RB">just</W>
        <W Lpos="IN">under</W>
        <ne id="ne314">
          <W Lpos="CD">ten</W>
        </ne>
      </unit>
    </sbar>
  </unit>
</s>
<s id="s121">
  <unit id="unit200">
    <W Lpos="RB">So</W>
    <W Lpos=":">-</W>
    <ne id="ne315">
      <W Lpos="NNP">XXX</W>
    </ne>
    <ve id="ve163" gId="nv163">
      <W Lpos="VBD">remarried</W>
    </ve>
  </unit>
</s>
<s id="s122">
  <unit id="unit201">
    <W Lpos="CC">And</W>
    <ne id="ne316">
      <W Lpos="PRP$">my</W>
      <W Lpos="NN">mother</W>
    </ne>
    <ve id="ve164" gId="nv164">
      <W Lpos="VBZ">calls</W>
    </ve>
    <ne id="ne317">
      <W Lpos="PRP$">her</W>
      <W Lpos="NNP">Doris</W>
    </ne>
    <W Lpos="RB">as</W>
    <W Lpos="RB">well</W>
  </unit>
</s>

```

Figure 4: The text after anonymizing ‘My grandpa Gaunting’ to ‘XXX’

of applications such as text summarization systems (Steinberger et al., 2005) and it is believed to be adequate for a preprocessing component. An anonymization tool based on an anaphoric / coreference resolver could potentially simplify the task of anonymization by eliminating the need to identify all possible forms used to mention a particular individual. The first evaluations of the anonymization module in its present form are currently underway. Future work will include building a converter from the Nite Object Model (NOM) to MAS-XML and viceversa, so that the module can be called directly from NXT, and evaluation with actual users.

## Acknowledgments

SQUAD is a demonstrator project funded under the ESRC Qualitative Data Archiving and Dissemination Scheme (QUADS). We wish to thank Louise Corti, Elizabeth Bishop, and Maria Milosavljevic for help and suggestions.

## 5. References

- K. Bontcheva, M. Dimitrov, D. Maynard, V. Tablan, and H. Cunningham. 2002. Shallow methods for named entity coreference resolution. In *Chances de rferences et rsolveurs d’anaphores, workshop TALN 2002*, Nancy, France.
- R. Mitkov. 1998. Robust pronoun resolution with limited knowledge. In *Proc. of the 18th COLING*, pages 869–875, Montreal.
- M. Poesio and M. A. Kabadjov. 2004. A general-purpose, off the shelf anaphoric resolver. In *Proc. of LREC*, pages 653–656, Lisbon, May.
- M. Poesio and R. Vieira. 1998. A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216, June. Also available as Research Paper CCS-RP-71, Centre for Cognitive Science, University of Edinburgh.
- M. Poesio, M. A. Kabadjov, R. Vieira, R. Goulart, and O. Uryupina. 2005. Does discourse-new detection help definite description resolution? In H. Bunt, J. Geertzen, and E. Thijse, editors, *Proc. of the 6th International Workshop on Computational Semantics (IWCS)*, pages 236–246, Tilburg, January.
- J. Steinberger, M. A. Kabadjov, and M. Poesio. 2005. Improving LSA-based summarization with anaphora resolution. In *Proc. of EMNLP*, Vancouver, October.