Building Hybrid Knowledge Representations from Text

Josef Meyer and Robert Dale
Language Technology Group
Division of Information and Communication Sciences
Macquarie University, Sydney, NSW 2109
Australia

E-mail: {jmeyer|rdale}@mri.mq.edu.au

Abstract

A significant obstacle to the development of intelligent natural language processing systems is the lack of rich knowledge bases containing representations of world knowledge. For experimental systems it is common practice to construct small knowledge bases by hand; however, this approach does not scale well to large systems. An alternative is to attempt to extract the desired information from existing knowledge sources intended for human consumption; however, attempts to construct broad-coverage knowledge bases using in-depth analysis have met with limited success. In this paper we present some work on an alternative approach that involves using shallow processing techniques to build a hybrid knowledge representation that stores information in a partially analysed form.

1. Introduction

A central problem for any intelligent language processing technology is the availability of rich representations of world knowledge. For many laboratory systems, researchers often respond to this problem by constructing a small knowledge base by hand, with the expectation that any scaled-up version of the technology would have access to a broad-coverage knowledge base sourced from somewhere else.

One can try to address this particular bottleneck by trying to use natural language processing to construct the requisite knowledge bases by automatically analysing existing texts; however, such attempts have had very limited success, at least in part because of the fundamental bootstrapping problems this raises. Such approaches have also generally underestimated the difficulties in achieving broad coverage even at the syntactic level; against this background, the aim of broad semantic and pragmatic coverage is unlikely to be satisfied for some considerable time to come.

In this paper, we take an alternative approach. Encouraged by the results achieved in shallow text processing in the information extraction community (see, for example, [10, 11]), we are interested in seeing how much usable knowledge we can extract from texts if we abandon the assumption that the end-result of such a process should be a knowledge base of the traditional kind. Our aim is to develop technologies that can build what we might think of as HYBRID KNOWLEDGE REPRESENTATIONS, where the representation used for information is only as deep as broad coverage analytic techniques will reliably permit. In other work [12], our group has shown that such representations can play a useful role in natural language generation applications, and we believe they may be useful in other language technology applications such as text summarisation.

In Section 2, we present more fully our notion of a hybrid knowledge representation, and argue for its usefulness. In Section 3, we describe some initial experiments in extracting hybrid knowledge representations from a corpus of real texts. In Section 4, we identify the specific problems that we have encountered so far, and discuss how these might be overcome. Finally, in Section 5 we draw some conclusions and point the way forward.

2. Hybrid Knowledge Representations

Traditional models of real world knowledge for natural language processing are built around the assumption that we can identify entities in the domain, and express symbolically relations between such entities and predicates over them. Figure 1 shows an extract of a simple knowledge base that takes this form.

Constructing a large knowledge base of this kind is extremely time consuming if carried out manually, and be-

```
(event e1
    (type buy) (buyer m1) (bought c1) (time t2))
(state s1
    (type expects) (time t1) (expecter j1) (event e1))
(< t1 t2)
```

Figure 1. Simple knowledge base entry for the phrase *John expects Mary to buy a cake***.**

yond the state of the art if we aim to do it automatically by analysing existing textual sources. One response to this problem has surfaced in the information extraction community: there, no pretense is made to provide a full and complete analysis of a given natural language text, but instead, only a very small set of very specific data elements is extracted. So, for example, in the domain of newswire reports about terrorist incidents, the key elements to be extracted might be the location of the event, some indication of the number and nature of the casualties, and the name of the perpetrating organisation. This information may represent only a small fraction of the content of the overall text, but for specific tasks such as categorisation this can already be of significant benefit.

There is a correlate of this kind of shallow processing in work on natural language generation, where the concern is not to produce some meaning representation given a text, but rather to create a text given some meaning representation. This area of activity also meets with a knowledge representation problem, since, in essence, there are many things we would like to have our natural language generation systems say that we simply do not yet know how to encode in some formal logical language. Furthermore, for many purposes we don't *need* to have particularly deep and sophisticated representations of such information. An approach adopted in many NLG systems, but first explicitly argued for in [12], is to use canned fragments of text with associated annotations that provide semantic tags that can be reasoned with. Figure 2 shows an extract from the Peba-II knowledge base, where properties predicated of entities are represented by means of the strings that would be used to realise these properties, along with simple annotations that allow the system's text planner to work out which information to convey. The usefulness of these representations is determined by the sophistication of the semantic tags attached to the text strings. In particular, simple annotations of the kind shown here rule out many kinds of inference that one might want to carry out: for example, they cannot support a question–answering system, such as that targeted in Hull and Gomez's [8] SNOWY system, which performs full syntactic and semantic analysis to build up a detailed knowledge base for topics relating to the dietary habits of animals. They are, however, adequate for at least one important class of applications: those focussed on the kind of tailored information provision task carried out by Peba-II and many other NLG applications.

We will refer to these representations as HYBRID REPRESENTATIONS since they combine traditional symbolic modelling—we have symbolic constants corresponding to the entities in the domain—with the use of text fragments for representing specific predications over these entities. If we have the technology required to decompose some of these text fragments into more symbolically-oriented representations, so much the better; the point is that we want to be able to express all the available information at some level of representation, rather than just the subset that we can reliably analyse at a deep level.

In the case of Peba-II, the hybrid knowledge base was built manually. Our goal is to see if we can build such representations automatically from text. In the next section, we describe our first experiments towards achieving this goal.

3. Extracting Information from a Corpus

3.1. An Overview of the Corpus

The corpus that serves as the basis for our experiments is based on two sets of entries drawn from those sections of two electronic encyclopedias (Microsoft Encarta and Grolier's Encyclopedia) that deal with animals. From this corpus, entries that consist of something other than a description of one or more species of animal have been excluded. The remaining entries comprise 803 entries from Encarta (87% of the original number), and 1107 from Grolier's (86% of the original number). There are also 603 titles that occur in both sets of entries; this 'parallel corpus' forms the basis for future work on integrating knowledge from different sources, and for carrying out analyses of variation between corpora.

An informal classification of the topics covered in ten pairs of entries with the same subject shows that there are a few topics that are commonly discussed, some in almost all entries. These include subjects such as the following (in roughly descending order of frequency): the classification of the animal within the linnaean system; its size, weight

```
;; Okapi
(ako Okapi Genus-Okapia)
(lex Okapi (sem ((type true-name))) (orth "Okapi")
   (syn ((category np) (agr ((number singular))))))
(hasprop Okapi (linnaean-classification Species))
(lex Okapi (sem ((type linnaean-name))) (orth "Okapia johnstoni")
   (syn ((category np) (agr ((number singular))))))
(hasprop Okapi (geography found-rainforest-zaire))
(lex found-rainforest-zaire (orth "is found in a small area of rainforest
   in Zaire") (syn ((category vp) (agr ((number singular))))))
(hasprop Okapi (height (quantity (lower-limit (unit m) (number 1.9))
   (upper-limit (unit m) (number 2)))))
(hasprop Okapi (weight (quantity (lower-limit (unit kg) (number 210))
   (upper-limit (unit kg) (number 250)))))
(hasprop Okapi (habitat lives-dense-forest))
(lex lives-dense-forest (orth "lives in dense forest")
   (syn ((category vp) (agr ((number singular))))))
(hasprop Okapi (diet eats-leaves-bark-shoots))
(lex eats-leaves-bark-shoots
   (orth "eats leaves, bark and shoots, some flowers, seeds and fruit")
   (syn ((category vp) (agr ((number singular))))))
```

Figure 2. Knowledge base entry from the Peba-II [12] system.

and colour; distinguishing features and their use (such as the claws and tongue on an Aardvark, which are used for feeding); information on diet and feeding habits; and information on reproduction such as the number of young in a litter.

3.2. Approach

The task that we are trying to achieve in constructing the hybrid knowledge base essentially consists of taking the text apart and replacing (or associating) certain noun phrases with symbolic referents; the material presented in the remainder of the sentences then corresponds to knowledge about these symbolic referents. Carrying out this partitioning of the raw material enables us (a) to assemble new texts using different subsets of information (possibly from multiple sources), and (b) to introduce the correct anaphoric relationships in the resulting texts. To produce appropriate results, we also need to tag the fragments of text with some indication of their semantics: this is needed both in order to to select which pieces of information should be included in a document, and to recognise when two sources are providing information on the same topic. For our system to be effective it needs to be able to perform this semantic tagging completely automatically. This can be achieved in a number of ways, for example by using the root form of the verb as a tag, perhaps in conjunction with any head nouns appearing in the complement; or, if we desire more general categories (as is likely), by using terms drawn from some classification over verbs as might be provided by a thesaurus-like resource such as WordNet.

In theory, our model is very simple. First, the system divides the text into sentences. Then, it identifies the subject NP, and replaces this NP with a corresponding symbolic referent, possibly one that already exists if we can infer an anaphoric relationship to an earlier mentioned entity. The rest of the sentence then becomes the complement that carries the relevant predication: the semantic annotation routine categorises the complement with one of a finite set of semantic tags, thus producing a hybrid knowledge base.

Of course, this simple model involves quite a number of assumptions that may not hold up. In particular:

- sentences do not all come in neat ⟨Subject VP⟩ active declarative form, with one major clause in each sentence; this complicates the process of separating out the entities which correspond to symbolic referents;
- without broad-coverage parsing technology, identifying the subject NPs even in simpler sentences is not always trivial; and
- identifying anaphoric relationships is not straightforward, as evidenced by the substantial body of work that continues to pursue a general solution to this problem.

Just how serious these problems are in the present case is still under investigation. In particular, we need to develop a more accurate picture of the types of sentence that occur in the corpus, and their frequency of occurrence. Our primary objective here is to determine the proportion of the sentences that can be processed in the manner described above without the need to resort to in-depth syntactic analysis. If this proportion is sufficiently high, then our approach is viable. A point worth noting is that we do not necessarily have to achieve complete or even nearly complete coverage: our hypothesis is that, given a sufficient number of textual sources, then all (or at least most) of the information will occur in a way that is easy to extract in at least one of the sources. Using the pairs of entries in our corpus that share the same topic provides a way of examining this and similar hypotheses.

3.2.1. Domain Dependent Solutions. The type of corpus that we have chosen shows consistency in both content (i.e. the general type of information that is expressed) and in style. This suggests that fairly simple statistical methods may prove useful in identifying features in the corpus to which we should pay additional attention.

We adopt the position that our techniques should be sufficiently domain-independent to be adaptable to similar but different corpora; however, we should take advantage of regularities in a specific corpus where they exist. To maintain portability we seek to explore generic techniques for leveraging the development of domain and corpus specific solutions. This is a technique that has been applied successfully in the information extraction community to develop robust systems that require minimal effort in porting to new domains [6].

An example of such a technique involves looking at frequently occurring n-grams to find key phrases, which can then be treated as special cases in subsequent processes applied to the text. This has already proved useful in diagnosing problems with part of speech tagging and automatically correcting certain classes of tagging error: looking at a list of bigram frequencies, we can note that the bigram *classified as* occurs 1003 times in the corpus, and that the following two words are incorrectly tagged by an automatic part-of-speech tagger in a significant number of cases. By tagging such common constructions correctly in a preprocessing step, deficiencies in the automatic tagger can be overcome.

Another example of where simple statistics gathered from the corpus have proved useful is in identifying the fact that certain noun phrases such as *the male(s)*, *the female(s)*, and *the animals* are amongst the most frequently occurring in the corpus. They also generally appear as a specific kind of associative reference whose antecedent is some animal type mentioned previously. A general solution to the resolution of such references would be difficult to construct; but their specific nature in our corpus, in conjunction with their frequency, warrants the construction of special case heuristics that perform anaphor resolution with a high degree of accuracy. Determining what sort of statistical analysis of the corpus is most likely to be useful, and automating as much of this analysis as is possible, is one key objective of our work.

3.3. Processing the Text

3.3.1. Preprocessing. The first step performed in processing the entries is to convert them to a consistent form in which the logical structure of the entry is preserved using an XML [2] based markup. Each entry is split into three sections: HEADER and FOOTER sections containing leading and trailing comments, and a BODY section containing the main text. Paragraphs are marked within the BODY section. In the second phase these paragraphs are split into sentences; simple heuristics prevent sentences from being split at initials and other abbreviations. Then, in the third and final phase, headings within the BODY section are recognised and re-tagged, and bylines are moved into a separate section.

The main reason for using a markup language is that it allows structure to be added to the document in subsequent stages of processing without destroying existing document structure. Additional advantages are that the use of markup is robust (in that it can survive changes to the document) and can be interpreted without special editing tools. There are two main reasons why XML in particular was chosen as a markup language: (a) standards are under development for annotating discourse relations in text using SGML and XML [5, 7]; and (b) there are freely available tools and libraries for processing XML documents.

After this initial preprocessing, a script tokenizes the sentences within the BODY section, and then passes them on to a part-of-speech tagger: we currently use Brill's [3], but almost any tagger could be plugged in with virtually no modification to the system.

3.3.2. Phrase recognition (chunking). The syntactic processing component of our system consists of a series (or *cascade*) of deterministic finite state transducers that each recognise some syntactic constituent. This is based on the principles described in Abney's [1] work on shallow parsing, and represents a simplified form of SRI's FASTUS [10] used in several of the DARPA's message understanding conferences. By focusing on the identification of phrase-sized chunks, we overcome some of the problems that face systems which attempt to derive full syntactic parses.

Our chunk parser currently recognises sequences of part-of-speech tags matching a regular expression and rewrites

them surrounded by XML tags appropriate for the type of chunk recognised; it does not re-write tags, and once an expression is recognised and tagged, the chunk is treated as a single word with a part of speech corresponding to its type. In its current form it is thus roughly equivalent to a deterministic bottom-up parser using a context-free grammar.[1] The next section describes the results of this mechanism in more detail.

3.4. A Worked Example

This section illustrates what the project is intended to achieve, and discusses the progress that we have made towards this end. Figure 3 shows the entry for the *okapi* from Grolier's encyclopedia after it has passed through our preprocessor, with everything but the BODY section removed for the sake of brevity. This is used to illustrate the type of representation that we intend to construct from the data (Fig. **??** and **??**) , and as a basis for discussing the progress that we have made towards this end. We then sketch a possible solution to the next component that we plan to implement, the anaphora resolution component. After this, we conclude by discussing some of the problems that are currently hindering progress on the work, our plans for overcoming them.

Noun phrases, particularly in the first few sentences dealing with a new topic, often incode information that we are interested in extracting. Ideally we would represent the subject of sentence 1 with a set of knowledge-base entries similar to the example from Peba-II shown in Figure 2[2]. Figure 4 shows a possible representation for the subject. Obtaining much of this information requires some knowledge of either the domain or the general structure of the type of text. For instance, identifying the fact that the second name in each of the appositive structures in sentence 1 is a scientific name, while the first is a common name, requires the use of some domain (or perhaps genre) specific heuristic. Identifying the use of *in* as a marker of type membership falls into a similar category.

Simple statements with a readily identifiable subject and predicate are highly ameniable to transformation into a form that can be used in the knowledge base, provided that any anaphoric reference within the statement can be resolved. The main task for these sentences involves assigning the predicate a semantic label that can be used for selecting relevant content in the script-based generation approach used in Peba-II and similar systems. For instance, from sentence (4) we could derive the following representation:

```
(related-to Okapi-male Okapi
   (reference-type type-subtype)
   (relation male-of-type))
(hasprop Okapi-male
   (special-feature has-small-horns))
(lex has-small-horns
   (orth "has small horns")
   (syn ((category vp))))
```

In this we have one knowledge base predicate encoding the semantic relationship between *males* and *okapi* that is realised as a type of *associative anaphora* [13, 9] in the text, and another that represents the core relationship that is expressed by the clause. The semantic tag *special-feature* can be determined easily from the verb group — finer grained distinctions such as distinctions between the posession of a physical characteristic (*have an opposing thumb*) and a property (*have a keen sense of smell*) may have to take into account the semantic types of the arguments to the verb.

As will be discussed later, recognising anaphoric relationships is crucial to the re-use of text fragments in documents that are generated 'on the fly'. Our knowledge representation extends that of Peba-II to allow for the representation of associative relationships by introducing a new predicate `related-to`, as shown below: In subsequently generated text, the term `Okapi-male` could be represented as *the male Okapi* or simply *the male* depending on whether there is a prior reference to Okapi that is accessible as an antecedent for the associative reference.

The Peba-II system stores certain types of property, such as height and weight, in a more abstract form than properties such as physical characteristics. Because these properties are so common, and generally easy to identify using shallow techniques, they can be treated as special cases when creating a knowledge base. For instance, the desired representation for sentence 3 would be the following:

```
(related-to Okapi-female Okapi
   (reference-type subtype)
   (relation female-of-type))
(hasprop Okapi-female (height (quantity
   (approximate (number 1.65) (unit m)))) (id 1))
(hasprop Okapi-female (length (quantity
   (approximate (number 1.65) (unit m)))) (id 2))
(link (general 1 2))
```

The specification of how the clauses are linked can be omitted in this case, but is important in the case where the clauses are linked by a subordinator such as *in that*, which appears in sentence 1.

Recognising when discourse relationships between fragments of text prevent one text fragment from being used

[1] A minor difference is that we allow finite lookahead in the patterns specified.

[2] Following the Peba-II example, classes of predicates such as those describing *size*, *habitat*, and *linnaean classification* are categorised according to an application-defined semantic hierarchy; these can be derived, for example, from the WordNet class of the main verb, as suggested in section 3.2.

```
⟨BODY⟩ ⟨P⟩
⟨S id="1"⟩The okapi, Okapia johnstoni, in the giraffe family, Giraffidae, is unusual
among mammals in that the female is larger than the male.⟨/S⟩ ⟨S id="2"⟩She may stand 1.65
m (5.5 ft) at the shoulder and be 1.95 m (6.5 ft) long.⟨/S⟩ ⟨S id="3"⟩The neck is tall,
the muzzle pointed, and the ears large and erect.⟨/S⟩ ⟨S id="4"⟩Males have small horns.⟨/S⟩
⟨S id="5"⟩The okapi can extend its long tongue to its eyes to wash them.⟨/S⟩ ⟨S id="6"⟩The
coat is purplish brown, with a light-colored face and bars of black and white on the
upper legs and buttocks.⟨/S⟩ ⟨S id="7"⟩Okapis live in dense eastern Congo rain forests.⟨/S⟩
⟨S id="8"⟩They are cud-chewers and eat fruits, leaves, and seeds.⟨/S⟩
⟨/P⟩ ⟨/BODY⟩
```

Figure 3. The entry for "Okapi" from Grolier's encyclopedia, as it appears after the first three stages of pre-processing. To save space, only the BODY element is shown.

```
(lex Okapi (sem ((type common-name))) (orth "Okapi")
   (syn ((category cn) (agr ((number singular))))))
(lex Okapi (sem ((type linnaean-name))) (orth "Okapia johnstoni")
   (syn ((category np) (agr ((number singular))))))
(lex Giraffe-family (sem ((type common-name) (level family)))
   (orth "Giraffe family") (syn ((category cn)) (agr ((number singular)))))
(lex Giraffe-family (sem ((type linnean-class) (level family))) (orth "Giraffidae")
   (syn ((category np)) (agr ((number singular)))))
(hasprop Okapi (belongs-to-family Giraffe-family))
```

Figure 4. Example of a knowledge base entry generated from the subject of sentence (1) in Fig. 3, "The okapi, Okapia johnstoni, in the Giraffe family, Giraffidae, . . . ".

(unmodified) without the other is a major concern when re-using text fragments; for this reason it is important thet the relationship between certain types of clause be encoded in the knowledge base. In sentence 1 the clause introduced by the complex subordinator *in that* justifies the main clause, and arguably also restricts its interpretation. This needs to be represented in the knowledge base, which can be done using the *link* predicate introduced above:

```
(hasprop Okapi-female (size female-larger-than-male)
   (id 3))
(hasprop Okapi-female (relation-to-supertype
   unusual-among-mammals) (id 4))
(link (justification 3 4))
```

Taking the main clause *The okapi . . . is unusal among mammals* as a paraphrase of the entire sentence would be at best marginally acceptable. In this case the second clause could be used in isolation, although not as a paraphrase of the sentence. Discourse relationships of this kind need not be explicitly signalled by the use of a cue-phrase like *in that*; however, this is usual in cases where clauses are particularly closely related.

The work that has been done in implementing a system to build a knowledge representation from a set of encyclopedia entries is still at an early stage. The preprocessing works well, but the tagging and chunk parsing are still unreliable. We have designed algorithms for resolving the types of associative reference that occur frequently in the corpus, but these are still to be implemented.

Figure 5 shows the structure produced by the chunk parser when it is run over the text from Fig. 3. As can be seen, there are still significant problems with the accuracy of both part of speech tagging and chunking. The errors in tagging can be to some extent explained by the fact that the corpus on which the tagger was trained is substantially different to that on which it is being used. The poor performance of the chunk parser is partially explained by the fact that the finite state transducers that drive it have been developed by hand in a fairly ad-hoc manner over a short period of time. We anticipate a considerable improvement in the quality of the output when enough of the system is developed to make traiing of these components feasible. This

```
⟨BODY⟩ ⟨P⟩ ⟨S id="1"⟩⟨N⟩The/DT okapi/NN⟨/N⟩ ,/, ⟨N⟩Okapia/NNP johnstoni/NN⟨/N⟩ ,/, ⟨PP⟩in/IN
⟨N⟩the/DT giraffe/NN family/NN⟨/N⟩⟨/PP⟩ ,/, ⟨N⟩Giraffidae/NNP⟨/N⟩ ,/, ⟨V1⟩is/VBZ⟨/V1⟩
unusual/JJ ⟨PP⟩among/IN ⟨N⟩mammals/NNS⟨/N⟩⟨/PP⟩ in/IN that/DT ⟨S2⟩⟨S1⟩⟨N⟩the/DT female/NN⟨/N⟩
⟨V1⟩is/VBZ⟨/V1⟩ larger/JJR ⟨PP⟩than/IN ⟨N⟩the/DT male/NN⟨/N⟩⟨/PP⟩⟨/S1⟩⟨/S2⟩ ./.  ⟨/S⟩
⟨S id="2"⟩⟨S2⟩⟨S1⟩⟨N⟩She/PRP⟨/N⟩ ⟨V1⟩may/MD stand/VB⟨/V1⟩ ⟨N⟩⟨N⟩1.65/CD m/NN⟨/N⟩ (/( ⟨N⟩5.5/CD
ft/NN⟨/N⟩ )/SYM⟨/N⟩ ⟨PP⟩at/IN ⟨N⟩the/DT shoulder/NN⟨/N⟩⟨/PP⟩⟨/S1⟩ and/CC ⟨V1⟩be/VB⟨/V1⟩
⟨N⟩⟨N⟩1.95/CD m/NN⟨/N⟩ (/( ⟨N⟩6.5/CD ft/NN⟨/N⟩ )/SYM⟨/N⟩⟨/S2⟩ long/JJ ./.  ⟨/S⟩
⟨S id="3"⟩⟨S2⟩⟨S1⟩⟨N⟩The/DT neck/NN⟨/N⟩ ⟨V1⟩is/VBZ⟨/V1⟩ tall/JJ⟨/S1⟩⟨/S2⟩ ,/, ⟨S2⟩⟨S1⟩⟨N⟩the/DT
muzzle/NN⟨/N⟩ ⟨V1⟩pointed/VBD⟨/V1⟩⟨/S1⟩⟨/S2⟩ ,/, and/CC ⟨N⟩the/DT ears/NNS⟨/N⟩ large/JJ and/CC
erect//JJ ./.  ⟨/S⟩ ⟨S id="4"⟩⟨S2⟩⟨S1⟩⟨N⟩Males/NNPS⟨/N⟩ ⟨V1⟩have/VBP⟨/V1⟩ ⟨N⟩small/JJ
horns/NNS⟨/N⟩⟨/S1⟩⟨/S2⟩ ./.  ⟨/S⟩ ⟨S id="5"⟩⟨S2⟩⟨S1⟩⟨N⟩The/DT okapi/NN⟨/N⟩ ⟨V1⟩can/MD
extend/VB⟨/V1⟩ ⟨N⟩its/PRP$ long/JJ tongue/NN⟨/N⟩ ⟨PP⟩to/TO ⟨N⟩its/PRP$ eyes/NNS⟨/N⟩⟨/PP⟩⟨/S1⟩⟨/S2⟩
to/TO ⟨V1⟩wash/VB⟨/V1⟩ ⟨N⟩them/PRP⟨/N⟩ ./.  ⟨/S⟩ ⟨S id="6"⟩⟨S2⟩⟨S1⟩⟨N⟩The/DT coat/NN⟨/N⟩
⟨V1⟩is/VBZ⟨/V1⟩ purplish/JJ⟨/S1⟩⟨/S2⟩ brown/NN ,/, ⟨PP⟩with/IN ⟨N⟩a/DT light-colored/JJ
face/NN and/CC bars/NNS⟨/N⟩⟨/PP⟩ of/IN black/JJ and/CC white/JJ ⟨PP⟩on/IN ⟨N⟩the/DT upper/JJ
legs/NNS and/CC buttocks/NNS⟨/N⟩⟨/PP⟩ ./.  ⟨/S⟩ ⟨S id="7"⟩⟨S2⟩⟨S1⟩⟨N⟩Okapis/NNP⟨/N⟩
⟨V1⟩live/VB⟨/V1⟩⟨/S1⟩⟨/S2⟩ ⟨PP⟩in/IN dense/JJ ⟨N⟩eastern/JJ Congo/NNP rain/NN
forests/NNS⟨/N⟩⟨/PP⟩ ./.  ⟨/S⟩ ⟨S id="8"⟩⟨S2⟩⟨S1⟩⟨N⟩They/PRP⟨/N⟩ ⟨V1⟩are/VBP⟨/V1⟩
⟨N⟩cud-chewers/NNS⟨/N⟩⟨/S1⟩ and/CC ⟨V1⟩eat/VB⟨/V1⟩ ⟨N⟩fruits/NNS⟨/N⟩⟨/S2⟩ ,/, ⟨V1⟩leaves/VBZ⟨/V1⟩
,/, and/CC ⟨N⟩seeds/NNS⟨/N⟩ ./.  ⟨/S⟩ ⟨/P⟩⟨/BODY⟩
```

Figure 5. The encyclopedia entry from Fig. 3 after chunking.

will not address all of the problems in tagging and chunk parsing, since there are some phenomena that are genuinely hard to deal with using this sort of technique.

The problems in assigning analyses to sentences 3, and 6 point to some basic inadequacies in the chunk parser and its 'grammar' of finite state transducers. Sentence 3 is problematic because of the verb elipsis in the second and third clauses; not only is this currently not handled by the chunk parser (which is relatively easy to fix), but because the local context for the second element resembles that for a verb, mistaggings are common. The problem with sentence 6 (apart from the mistaggings) is related to the use of conjunctions. Conjuctions tend to produce local (and occasioanlly global) ambiguities that affect even conventional parsers. However, a deterministic parser does not have the option of backtraching once it has started to pursue an incorrect path, so local ambiguities pose more of a problem. The use of embedded clauses also poses a potential problem for cascaded finite state transducers, but examples like that in sentence 5 are simple enough to handle as a special case, and common enough to be worth treating in this way.

One of the goals in this project is to develop a set of tools to assist in the development of the language models used in the different stages of processing. One of the main objectives is to produce (or collect) tools for annotating the corpus at a range of levels from part of speech tagging and chunking to the representation of anaphoric and inter-clausal relationships.

Despite the inaccuracy of the tagger, the system is usually reasonably successful in picking out base noun phrases (that is, noun phrases without postmodification). Identifying noun phrases (or at least base noun phrases) is probably the most important task of the syntactic analysis component because it is a prerequisite for anaphora resolution. The presence of more structure allows for a finer grained decomposition of the texts and provides more information that can be used to improve the accuracy of later stages of analysis. However, as shown by recent work on the resolution of pronominal anaphora [4], it is possible to get remarkably good results with little more than position, a bit of lexical semantics, and part of speech information.

4. Conclusions

In this paper we have presented an approach to knowledge base construction from encyclopedic texts that results in a hybrid knowledge base that stores data at different levels of linguistic realisation. This type of knowledge base has been shown to be useful in natural language generation systems that generate web-pages where the content is tailored to the user's needs, and we anticipate that it may also prove useful in such areas as text summarisation. The purpose of the work described in this paper is to determine if it is feasible to employ shallow processing techniques at a number of levels to construct such a knowledge base.

We have contrasted our approach with previous work by Gomez, Hull, and Segami [8] that uses more in-depth syntactic and semantic analysis to produce a knowledge base

that can be used in a question answering system from a similar corpus; our argument is that the problem that we are trying to solve is actually significantly different—we are aiming to produce a knowledge base with greater coverage for applications where a fine-grained knowledge representation is neither necessary nor desirable.

The current focus of our work is on the natural language processing aspects of the task, rather than the issues concerning knowledge base construction. The main novel aspect of the research in which we are currently engaged centres of the resolution of certain kinds of associative anaphoric relationship that occur commonly in this type of descriptive text. There are also problems that need to be addressed in adapting existing components to a new domain so that it makes best use of domain-specific features without sacrificing too much portability. We have encountered practical problems obtaining accurate results from the part of speech tagger and the chunk parser. These will have to be addressed before we can produce results derived from the analysis of any significant amount of data.

For example, the frequency of certain classes of noun phrases, such as terms for body parts, and subtypes such as *male* and *female*, occur with high frequency, as do the pronouns *it* and *they*. This can be taken as an indication that coreference and forms of associative reference involving type–subtype and whole–part relationships are worth treating specially when analysing this corpus.

Although our experiment is still in its early stages, our preliminary results suggest that the broad-but-shallow approach we are adopting is capable of providing the kinds of knowledge representation we set out to build.

References

[1] Steven Abney. Partial parsing via finite-state cascades. In *Proceedings of the ESSLLI '96 Robust Parsing Workshop*, 1996.

[2] Tim Bray. *The Annotated XML 1.0 Specification*. September 1988. Available online as `http://www.xml.com/xml/pub/axml/axmlintro.html`.

[3] E. Brill. A simple rule-based part of speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, Trento, Italy, 1992.

[4] Branimir Boguraev and Christopher Kennedy. Anaphora in a wider context: Tracking discourse referents. In *Proceedings of the 12th European Conference on Artificial Intelligence*, 1996. Available as `http://www.ling.nwu.edu/~kennedy/Docs/kb-ecai96.ps`.

[5] N. Chinchor and L. Hirschman MUC-7 Coreference Task Definition, Version 3.0 In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, 1997. Available online as `http://www.muc.saic.com/proceedings/co_task.html`

[6] The PLUM System Group. BBN: Description of the PLUM system as used for MUC-5. In *Proceedings of the Fifth Message Understanding Conference (MUC-5)*, 1993.

[7] Sarah Davies, Massimo Poesio, Florence Bruneseaux, and Laurent Romary. Annotating coreference in dialogues: proposal for a scheme for MATE. First draft, available as `http://www.cogsci.ed.ac.uk/~poesio/MATE/anno_manual.html`, July 1998.

[8] Fernando Gomez, Richard Hull, and Carlos Segami. Acquiring Knowledge from Encyclopedic Texts. In *Proceedings of the 4th ACL Conference on Applied Natural Language Processing*, 1994.

[9] J.A. Hawkins. *Definiteness and Indefiniteness: a study of reference and grammaticality prediction.* London: Croom Helm. 1978.

[10] Jerry R. Hobbs, Douglas Appelt, John Bear, David Israel, Megumi Kameyama, and Mabry Tyson. SRI International: Description of the JV-FASTUS System Used for MUC-5. In *Proceedings of the Fifth Message Understanding Conference (MUC-5)*, pages 221–235, Baltimore, Maryland, August 25–27 1993.

[11] W. Lehnert, J. McCarthy, S. Soderland, E. Riloff, C. Cardie, J. Peterson, F. Feng, C. Dolan, and S. Goldman. UMass/Hughes: Description of the Circus System Used for MUC-5. In *Proceedings of the Fifth Message Understanding Conference (MUC-5)*, pages 221–235, Baltimore, Maryland, August 25–27 1993.

[12] Maria Milosavljevic, Adrian Tulloch, and Robert Dale. Text generation in a dynamic hypertext environment. In *Proceedings of the Nineteenth Australasian Computer Science Conference (ACSC'96)*, pages 417–426, Melbourne, Australia, 31 January – 2 February 1996.

[13] Renata Vieira. *Definite description processing in unrestricted text*. PhD thesis, University of Edinburgh, 1998.