eman ta zabal zazu

Universidad del País Vasco

Euskal Herriko Unibertsitatea

Automatic anaphora resolution in Kyoto project

Sonia Ortiz de Arri
Tutorea: Arantza Diaz de Ilarraza

hap

Hizkuntzaren Azterketa eta Prozesamendua Masterreko titulua
lortzeko bukaerako proiektua

2012eko iraila

Sailak: Lengoaia eta Sistema Informatikoak, Konputagailuen Arkitektura eta Teknologia, Konputazio Zientziak eta Adimen Artifiziala, Euskal Filologia**,** Elektronika eta Telekomunikazioak.

Laburpena

Lan honen helburua ingelerazko korreferentzien ebazpenerako tresna modular bat Kyoto proiektuan integratzea izan da. Kyoto proiektuaren helburuetako bat semantikan oinarritutako informazio-erauzketa sistema eleaniztuna sortzea da eta helburu horri lotuta, beste lan askoren artean, korreferentzien ebazpen automatikoa egiten duten tresnen erabilera eta integrazioa egitea aurreikusi da. Lan honetan, korreferentzien ebazpen automatikoa egiten duten hainbat tresna ebaluatu ditugu: GuiTAR, BART, Reconcile, SUCRE eta Stanford-eko CoreNLP. Haien azterketa egin ondoren, proiektuaren beharretara hobekien egokitzen zena BART zela erabaki eta tresna hori Kyotoko gainontzeko tresna linguistikoen artean integratzeko urratsak egin ditugu: batetik, beharrezkoa izan da BART tresnarentzat egokitze modulu bat diseinatu eta implementatzea, eta bestetik, integratutako tresna horrek korreferentzien inguruan eskaintzen digun informazio semantikoa Kyoto proiektuko KAF fitxategietan behar bezala txertatu behar izan dugu. Lan honen azken helburua, Kyotoko informazio erauzketarako errobotek (kybot-ek) baliatzen duten informazio semantikoa aberastea izan da.

Hitz gakoak: informazio-erauzketa, korreferentzien erresoluzioa, anafora, informazio semantikoa, Kyoto proiektua, kybot.

Abstract

The aim of this work has been an approach to integrate a modular toolkit for coreference resolution system for English language in Kyoto project, a multilingual semantic based information extraction system. To achieve this objective we have analyzed some tools that try to resolve the coreference problem in some way or another: Guitar, BART, Reconcile, SUCRE, and Stanford's CoreNLP. We have made an evaluation and we have selected the one that better fits our needs. For the integration of the selected modular toolkit for coreference resolution in our project we have designed and developed a system that preprocesses the output of such toolkit and interprets the given coreferences and its antecedents. Finally, the developed system includes that semantic information about the coreferences in Kyoto project's KAF file (along with the rest of the morphosyntactic information) and this way, computer programs named kybots will be able to detect more and better concept instances and relations in the texts.

Keywords: information extraction, coreference resolution, anaphor, semantic information, Kyoto project, kybot.

Table of Contents

1.- Introduction

Coreference resolution is the process of detecting noun phrases in a document and determining whether these noun phrases refer to the same entity or not. As defined in ACE (2005), an entity is “an object or set of objects in the world”. Practical applications in natural language processing (NLP) of coreference resolution are summarization, Question Answering and Information Extraction (IE) since they all can take advantage of the identification of coreference relations between noun phrases in general.

The explosive growth in the last decade of electronic textual information available on private networks and on the Internet has brought the need to handle in an automatic way that huge amount of information to be able to obtain the data we require at each moment. The goal of Information Extraction is to extract from documents (which may be in a variety of languages) salient facts about previously specified types of events, entities or relationships and coreference resolution can improve substantially the final results of this crucial task.

Kyoto is an Information Extraction system offering knowledge transition and information across different target groups, transgressing linguistic, cultural and geographic boundaries. Initially developed for the environmental domain, Kyoto will be usable in any knowledge domain for mining, organizing, and distributing information on a global scale. The output of the linguistic analysis of Kyoto project is stored in an XML annotation format called Kyoto Annotation Format KAF, (Bosma et al., 2009) and it stores the analyzed text and its morphosintactic

information. Kybots are computer programs that use the mined concepts and the generic concepts already connected to the language wordnets and Kyoto ontology to detect concept instances and relations in text (i.e. tropical species decreased by 15% last year). Kybots operate on KAF files. They use profiles that define patterns in KAF. If there is a match, they generate an output structure that can be added to the KAF file or outputted separately. The profiles consist of a list of variables, relations between the variables and an output structure.

The present work is an attempt to integrate a modular toolkit for coreference resolution into Kyoto project. To achieve this objective we have analyzed different modular toolkits for coreference resolution and we have chosen BART as the best that fits our needs. For the integration of BART in our project we have designed and developed a system that preprocesses the coreference information of BART, interprets it and converts it into a given format of antecedent-anaphor pairs to be used by kybots. This way, Kyoto project in general and kybots in particular will be able to detect more and better concept instances and relations in the texts.

The rest of the paper is laid out as follows: Section 2 gives an overview of the research literature related to this work. Section 3 describes the general objectives of our work and it resumes the steps taken to achieve our aim. Section 4 contains a summary of the coreference resolution systems that we have analyzed and explains in depth why we have chosen BART as the best system to integrate it in our project. Section 5 shows the architecture and implementation of our work. Section 6 gives the results of our evaluation for BART and finally Section 7 points out the conclusions and proposes future work to do.

2.- Related work

In this section we present some of the research literature related to coreference resolution and information extraction and we briefly describe a system that has Kyoto's similar objectives.

Noun phrase (NP) coreference resolution is an important subtask in NLP applications such as text summarization, information extraction, data mining, and question answering. In particular, information extraction systems have revealed that coreference resolution is such a critical component of IE systems that a separate coreference subtask has been defined and evaluated in the MUC (Message Understanding Conferences) and ACE (Automatic Content Extraction) programs since 1995.

Two main approaches for Information Extraction have been proposed during the last years. One is based on knowledge-engineering and the other is a statistical or machine learning approach. In the knowledge based approach the expert skills play a crucial role in the successful identification and analysis of relevant information and even though the amount of manual work is considerable and the development and test cycle of the tagging is expensive, the results are very precise. The most representative examples of this kind of systems are ANNIE (Cunningham et al., 2000) that relies on finite state algorithms; KIM tool (Popov et al., 2003) that provides semantic search interface that hides complex query syntax; AquaLog (Lopez et al., 2005) that uses a controlled

language for querying the ontology; or the KnowItAll system (Etzioni et al., 2005) that took the next step in automating IE by learning to label its own training examples using only a small set of domain-independent extraction patterns.

The other approach in contrast, is based on machine-learning. This approach exploits artificial intelligence techniques to automatically induce from a Corpus, extraction rules starting from a set of generic information patterns. With this approach there is no need to develop expert written expensive grammars and although the results are not so precise, the process has a lower cost and still, the precision is reasonably good. Some of most recent examples of this kind of systems are ReVerb (Fader et al., 2011) that implements a novel relation phrase identifier based on generic syntactic and lexical constraints; R2A2 (Christensen et al., 2011) that adds an argument identifier to better extract the arguments for these relation phrases; TextRunner (Banko et al., 2007) that extracts high-quality information from sentences in a scalable and general manner; WOE (Wu and Weld, 2010) that uses dependency features (WOEparse) and training data generated using Wikipedia infoboxes to learn a series of open extractors (WOEpos); or Preemptive (Shinyama and Sekine, 2006) that avoiding relation-specificity, does not emphasize Web scalability.

Something similar happens with the coreference resolution task as well. A great deal of research has been done on this subtask lately, using approaches ranging from those based on linguistics to those based on machine learning.

The knowledge-based approach has the advantage in that, usually, little or no annotated corpora are required because an expert in linguistics provides the rules for filtering features for NP resolution. However, it does rely heavily on hand-crafted heuristics or rules, which also require large investments of time and effort to create. Zhou and Su (2006) presented a constraint-based multi-agent strategy. This strategy first uses general heuristics such as morphological and semantic consistency to filter out invalid antecedent candidates, and then an antecedent for the anaphor is chosen based on the principle of proximity. Bean and Riloff (2004) pioneered an approach to identify NP coreferences by using information extraction patterns to identify contextual role knowledge. This approach first identifies definite, non-anaphoric noun phrases, and then uses case resolution to identify the most easily resolved phrases.

In contrast, the machine-learning approach is corpus based. It requires a relatively small corpus of training documents that have been annotated with coreference information. Nevertheless, the bigger is the size of the corpus, the better will be the learning process. All possible markables in a training document are determined and training examples are generated for appropriate pairs of markables. These training examples are then given to a learning algorithm to build a classifier. To determine the coreference chains in a new document, all markables are determined and potential pairs of coreferring markables are presented to the classifier, which decides whether the two markables actually corefer or not. Soon et al. (2001) proposed a 12-feature classifier based on a decision tree, which returns a number between 0 and 1 to indicate the likelihood that two noun phrases corefer. Their training data came from and was applied to the MUC corpora. Ng and Cardie (2002a) extended this approach with three extra-linguistic changes: the clustering approach, the creation of training instances, and the definition of string match features. Yang et al. (2008) adopted a twin-candidate model for coreference resolution, which considered that the

purpose of classification was to determine the preference between two competing candidates for the antecedent of a given anaphor. Haghighi and Klein (2009) presented a system that was deterministic and was driven entirely by syntactic and semantic compatibility as learned from a large, unlabeled corpus. Haghighi and Klein (2010) used a generative model that exploited a large inventory of distributional entity types.

In the last years machine-learning approaches have spread widely and indeed this can be confirmed by the fact that the majority of the systems tested in SemEval-2010 and CoNLL-2011 were of this type.

As well as we intend in Kyoto project, some other systems have combined different approaches to create knowledge repositories across languages and they too have tried to improve their results by including coreference resolution systems as we have done with BART in Kyoto. Especially in the biomedical domain, systems have been developed that use rich knowledge resources such as bio-medical thesauri together with ontologies to detect data and facts with high precision in large document collections. The Bootstrapping Of Ontologies and Terminologies STrategic REsearch Project (BOOTStrep) is a good example of a project that combines resources and objectives similar to those of Kyoto project. Bootstrep learns the terminology from text and represents the results in an ontology. Then, the terminology and the ontology are used to apply text mining to document collection. It also uses a coreference resolution system based on an expressive entity-mention model that performs coreference resolution at an entity level (Yang et al. 2008).

3.- Coreference resolution

Several formal evaluations have been conducted for the coreference resolution task (MUC-6, 1995, ACE NIST 2004), and the data sets created for these evaluations have become standard benchmarks in the field (MUC and ACE data sets). However, it is still frustratingly difficult to compare results across different coreference resolution systems because reported coreference resolution scores vary wildly across data sets, evaluation metrics, and system configurations.

Most methods of coreference resolution, if providing a baseline, usually use a feature set similar to Soon et al. (2001) or Ng and Cardie (2002b) but a considerable engineering effort is needed to implement a complete end-to-end coreference resolution system. It is a complex problem, and successful systems must tackle a variety of non-trivial subproblems that are central to the coreference task and that require substantial implementation efforts. As a result, many researchers exploit gold-standard annotations, when available, as a substitute for component technologies to solve these subproblems. For example, many published research results use gold standard annotations to identify NPs, to distinguish anaphoric NPs from non- anaphoric NPs, to identify named entities, and to identify the semantic types of NPs. Unfortunately, the use of gold standard annotations for key/critical component technologies leads to an unrealistic evaluation setting, and makes it impossible to directly compare results against coreference resolvers that solve all of these subproblems from scratch.

Comparison of coreference resolvers is further made difficult by the use of several competing (and non-trivial) evaluation measures, and data sets that have substantially different task definitions and annotation formats. Additionally, coreference resolution is a pervasive problem in NLP and many NLP applications could benefit from an effective coreference resolver that can be easily configured and customized.

Central to the development of efficient and reliable approaches to automatic NP coreference resolution is the issue of what features should be used to identify the coreference. Ng and Cardie (2002b) listed 53 features, including gender agreement, number agreement, head noun matches, semantic class agreement, positional information, contextual information, apposition, abbreviation, and others. At one extreme, efficiency alone forbids the use of all of these features; at the other, no single linguistic feature is completely reliable. The most desirable features for use in coreference resolution are robust and inexpensive, perform well over various domains, and can be obtained automatically. Features may be lexical, grammatical, semantic, syntactic, contextual, or heuristic. Given the broad range of features that may be chosen, there is currently no definitive classification of their relative merits or effects on system performance.

Being Kyoto project a semantic based information extraction system for knowledge sharing, it should be a 'must to' to consider the coreference resolution problem a main issue to take into account to achieve the general objectives of the project. The coreference resolution might be in some cases the key that gives the hint when leading with semantic information and might help to answer questions and gather specific information that otherwise would stay hidden. We stated that giving to the whole information extraction system more accurate and richer semantic information, would have its results improved. Specifically, we thought that substituting pronouns or coreference elements in general, with their corresponding antecedents, would help to better extract knowledge from unstructured texts and that this could be done using some kind of coreference resolution system.

In this sense, the aim of this work is an approach to integrate a coreference resolution toolkit with a modular arquitecture into Kyoto project. This way, we will go beyond the handicaps mentioned above and it will be possible to take advantage of a coreference resolver that is easily configurable and customizable and that can be evaluated and compared using different metrics.

To succeed in our aim we carried out a research to find public modular coreference resolvers in English. We analyzed their behavior and evaluated them to see which one of them adjusted better to our needs. Next, we designed and developed a system that treated the output of the selected coreference resolver to adjust it to Kyoto specifications. Finally, we integrated the resulting semantic information in KAF files, for kybots in particular and Kyoto in general to take advantage of it.

4.- Modular architectures for coreference resolution

We will introduce in the next paragraphs a brief analysis of several architectures that have been

created for coreference resolution and are currently publicly available. They all try to address the issues mentioned in the previous section and can serve as a modular software infrastructure or platform to support the creation of, experimentation with and evaluation of coreference resolvers.

GuiTAR (Poesio and Kabadjov, 2004; Kabadjov, 2007) is one of the first anaphora resolution systems designed to be modular and usable as an off-the-shelf component of a NL processing pipeline. It was designed to be as independent as possible from the specifics of the modules used to extract information from – e.g., POS-taggers and parsers – and to be as modular as possible, allowing for the possibility of replacing specific components (e.g., the pronoun resolution component) and this way, it laid the foundations for the next generations of researchers dedicated to develop coreference resolution architectures.

BART (Versley et al., 2008) is based on code and ideas from the system of Ponzetto and Strube (2006), but also includes some ideas from GUITAR and other coreference systems (Versley, 2006; Yang et al., 2006). The goal of bringing together state-of-the-art approaches to different aspects of coreference resolution, including specialized preprocessing and syntax-based features has led to a design that is very modular.

Reconcile (Stoyanov et al., 2010) is a general architecture for coreference resolution that can be used to easily create various coreference resolvers and includes a comprehensive set of features that draw on the expertise of state-of-the-art supervised learning approaches, such as Bengtson and Roth (2008).

SUCRE (Kobdani et al., 2010) has a novel approach to model an unstructured text corpus in a structured framework by using a relational database model and a regular feature definition language to define and extract the features and it is language independent.

Stanford's CoreNLP (2011) provides a set of natural language analysis tools which can take raw English language text input and give the base forms of words, their parts of speech, whether they are names of companies, people, etc., normalize dates, times, and numeric quantities, and mark up the structure of sentences in terms of phrases and word dependencies, and indicate which noun phrases refer to the same entities.

4.1.- BART: Beautiful Anaphora Resolution Toolkit

BART (Versley et al., 2008) has been developed as a tool to explore the integration of knowledge-rich features into a coreference system. It is based on code and ideas from the system of Ponzetto and Strube (2006) and some other ideas from GUITAR and other coreference systems (Versley, 2006; Yang et al., 2006) and for this reason it has led to a design that is very modular. This design provides an effective separation of concerns across several tasks and roles that makes it possible to effortlessly combine functionality improvements created by independent efforts, including engineering new features that exploit different sources of

knowledge, designing improved or specialized preprocessing methods, and improving the way that coreference resolution is mapped to a machine learning problem. It also makes it very easy to explore possible configurations of these components to adapt to various accuracy and speed trade offs.

One goal for BART's architecture has been to provide effective separation of concerns for the following three groups of people who might be interested in working on a system for coreference resolution:

- Those who aim to do feature engineering, creating new features that exploit different sources of knowledge.
- Those who aim to explore different preprocessing methods, improving the quality of the input to coreference resolution proper.
- Those who aim to explore different methods of representing coreference resolution as a learning problem.

To reach this goal, there is a clean separation between the domains of preprocessing, feature extraction, learning and training/testing that we will explain as follows.

Preprocessing

Preprocessing consist in marking up noun chunks and named entities, as well as additional information such as part-of-speech tags and merging these information into markables that are the starting point for the mentions used by the coreference resolution proper. There are different ways to carry out this preprocessing:

- Using chunking pipeline:
  - It uses a classical combination of tagger and chunker, with the Stanford POS tagger (Toutanova et al., 2003), the YamCha chunker (Kudoh and Matsumoto, 2000) and the Stanford Named Entity Recognizer (Finkel et al., 2005).

- Using a parsing pipeline:
  - It supports Charniak and Johnson's reranking parser (Charniak and Johnson, 2005) to assign POS tags and uses base Nps as chunk equivalents, while also providing syntactic trees that can be used by feature extractors.
  - It also supports Berkeley parser (Petrov et al., 2006).

- Using the Carafe pipeline:
  - It uses an ACE mention tagger to provide a better starting point for mention detection on the ACE corpora.

Feature extraction

Feature extraction goes through all mentions and looks for possible coreferent elements in previous mentions. Those coreference sets are enriched with classification features by feature

extractors, and then handed over to a machine learning-based classifier that decides, given the features, whether the elements are coreferent or not. The set of feature extractors that the system uses is set in an XML description file. The default resolver looks for the basic features as described by Soon et al. (2001) but there are as well other features that can be configured.

- Soon et al. (2001) basic features:

  - *MentionType / MentionTypeBuggy*
    The features extractors *MentionTypeBuggy* and *MentionType* extract information about the form of the anaphor (definiteness, demonstrative, pronoun), the antecedent and also includes a feature that indicates whether the two mentions are both proper names.
    FE_MentionType_Buggy checks for the prefix “the” on the mention string.

  - *Gender*
    The feature extractor *Gender* uses gender information from the mention to assess gender compatibility. The assigned value can either be true, false, or unknown.

  - *Number*
    The feature extractor *Number* uses number information to determine number compatibility. This is either true or false.

  - *Alias*
    Alias uses the techniques described in (Soon et al., 2001) to match abbreviations and name variations.

  - *Appositive*
    Appositive adds a feature that is true whenever two mentions are separated exactly by a comma.

  - *String Matching*
    StringMatch strips the determiners off the markable string and then performs a case-insensitive comparison of the rest.

  - *SemanticClass*
    FE_SemanticClass uses the *SemanticClass* property of the mention to assess the semantic compatibility of anaphor and antecedent (either TRUE, FALSE, or UNKNOWN if either of the two has an unknown semantic class and the lexical heads do not match).

  - *SentenceDistance / DistDiscrete*
    SentenceDistance gives the distance of anaphor and antecedent candidate in sentences. DistDiscrete is meant as a discretisation of the values, with two binary features that indicate whether the candidate is in the same sentence or in the previous sentence.

- Syntactic features:

  - *SynPos*

    SynPos yields a string that is composed of the first three unique labels of parent nodes. This is meant to indicate the syntactic position. Subjects will have a value of ‘np.s’, whereas direct objects will have a value of ‘np.vp.s’, and a noun phrase embedded in a noun-modifying PP would have a value of ‘np.pp.np’.

  - *TreeFeature*

    The feature *TreeFeature* is a tree-valued feature that carries information about the syntactic relationship between anaphor and candidate. Its value is a subtree of a parse tree covering both the anaphor and the antecedent candidate. It includes the nodes occurring in the shortest path connecting the pronoun and the candidate, via the nearest commonly dominating node. Also it includes the first-level children of the nodes in the path.

- Knowledge-based features:

  - *WebPatterns*

    The *WebPatterns* feature extractor uses pattern search on the World Wide Web to find instance relations as they exist between ‘China’ and ’country’, or ‘Clinton’ and ‘president’. Queries are cached in a local BerkeleyDB-JE database.

  - *WikiAlias*

    The *WikiAlias* feature extractor uses information extracted from Wikipedia, namely redirects and links to a given page, but also appearance in lists, to provide evidence for name variations .

  - *Wiki*

    The *Wiki* feature extractor uses redirects and the category graph of Wikipedia to assess candidate relatedness, as described in (Ponzetto and Strube, 2006).

  - *WNSimilarity*

    The *WNSimilarity* feature extractor extracts the WordNet distance between antecedent and candidate heads, according to several distance measures.

  - *SemClassValue*

    The SemClass feature extracts the semantic class values of anaphor and antecedent, both alone and as a pair.

Learning

Learning is based in a generic abstraction layer that maps application-internal representations to a suitable format for several machine-learning toolkits.

- All classifiers of the WEKA machine learning toolkit (Witten and Frank, 2005).

  The weka classifier uses the WEKA machine learning toolkit for classification; all classifiers from WEKA can be used, and the class name of the corresponding classifier has to be given in the “learner” attribute. Options, as they appear on the command line shown by the WEKA Experimenter, can be specified in the “options” attribute.

- SVMLight / SVMlightTK variant which handles tree-valued features (Moschitti, 2006).

  The svmlight classifier uses SVMLight, either in its plain variant or in the SVMLight/TK variant. Options to svm learn can be specified in the “options” attribute.

- A Maximum Entropy classifier.

  The maxent classifier is a maximum entropy classifier built upon the LBFGS[1] implementation of Mallet[2]. It is able to perform feature combinations. Binary feature combinations give you a similar accuracy to the SVMLight polynomial-degree-2 classifier, with much reduced training times.

Training / testing

Training and testing is factored out into the encoder/decoder component, which is separate from feature extraction and machine-learning itself. This way, it is possible to completely change the basic behavior of the coreference system by providing new encoders/decoders, and still rely on the surrounding infrastructure for feature extraction and machine-learning components.

As we have shown, BART has an extremely modular and adaptable architecture that gives us the possibility to try a wide range of different preprocessing, features and classifiers. This reason and the fact that it is available as open source has led us to point out BART as the best choice to try our aim and so, we have selected it to try to integrate a coreference resolution system in our project.

4.1.1 BART Settings

It is possible to change BART options modifying some XML and Java files. For example, training and test set locations, preprocessing pipelines, features sets or machine learning options can be modified in the config.properties XML file. To change individual features or preprocessing options, it is necessary to comment the Java code.

We have used for our experiments the standard configuration that comes within the installation of BART. That is, to run a basic system, we have used the following external components:

1 http://users.eecs.northwestern.edu/~nocedal/lbfgs.html
2 http://mallet.cs.umass.edu/

Preprocessing:

- The YamCha chunker and the YamCha model collection (for the chunker-based pipeline)

Feature Extraction:

- The default resolver looks for the basic features (Soon et al. 2001):
  - MentionType and MentionTypeBuggy
  - Gender
  - Number
  - Alias
  - Appositive
  - String Matching
  - SemanticClass
  - SentenceDistance / DistDiscrete

Learning:

- SVMLight / SVMlightTK variant which handles tree-valued features (Moschitti, 2006).

5.- Architecture and implementation

As we already said, we have chosen BART (Versley et al. 2008), a modular toolkit for coreference resolution that supports state-of-the-art statistical approaches to the task and enables efficient feature engineering as the baseline to integrate a coreference resolution system into Kyoto project.

In the next subtasks we give the details of the architecture and implementation of our system.

5.1.- Architecture

Later in next subsections we will give the details and explain in depth the different subtasks that have been carried out for the implementation of our system. Now, we give first a summary of the pipeline steps:

Firstly, a KAF file of Kyoto project is taken as input.

1.- *Words Extractor:* Extracts from the KAF file only the raw text to be analyzed.

2.- *BART Coreference Resolution Toolkit:* Resolves for the raw text the existing coreference relations.

3.- *Token mismatches treatment:* Possible token mismatches between BART and KAF are treated.

4.- *Terms conversion:* Proposed coreference elements are converted to canonical forms.

5.- *Antecedent-Anaphor pairs conversion:* Canonical coreference elements are converted to antecedent-anaphor pairs.

6.- *Antecedent-Anaphor pairs integration in KAF file.*

As a result of this process, an enriched KAF file with anaphoric information is given as output.

5.2.- Implementation

Now, in the following lines we will go in depth into the subtasks listed above and we will try to clarify the details inherent to them.

5.2.1.- Words Extractor

The output of the linguistic analysis of Kyoto project is stored in an XML annotation format called Kyoto Annotation Format (KAF). This format incorporates standardized proposals for the linguistic annotation of text but represents them in an easy to use layered structure. In this structure, words, terms, constituents and syntactic dependencies of a text are stored separately with references across the structures. Since BART requires raw text to process its words and identify coreferences, we needed to extract the words listed in the KAF file to give them to BART as input. With this aim, we developed a words extractor in perl scripting language that uses the Lib::LibXML perl module and goes through the XML structured original KAF file to collect the words founded and give them as output in a separate file (*Words_File1)*.

In the next figures we will follow the processing of the sentence *"The Sub-Group chairmen and their staff have gone about their work with determination, enthusiasm and energy and I take this opportunity to thank them warmly for their efforts".*

Figure 5.1. corresponds to an original KAF file from Kyoto. This file will be the input of *Words Extractor*. (Note: For simplicity we have cut parts of the file. Refer to appendix to see the complete file.)

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<KAF xml:lang="en" doc="KyotoKAFEng_3/1286">
<text>
  <wf wid="w286" sent="15" page="1">The</wf>
  <wf wid="w287" sent="15" page="1">Sub-Group</wf>
  <wf wid="w288" sent="15" page="1">chairmen</wf>
  <wf wid="w289" sent="15" page="1">and</wf>
  <wf wid="w290" sent="15" page="1">their</wf>

…

  <wf wid="w310" sent="15" page="1">them</wf>
  <wf wid="w311" sent="15" page="1">warmly</wf>
  <wf wid="w312" sent="15" page="1">for</wf>
  <wf wid="w313" sent="15" page="1">their</wf>
  <wf wid="w314" sent="15" page="1">efforts</wf>
```

```
  <wf wid="w315" sent="15" page="1">.</wf>
</text>
<terms>
 <term tid="t204" type="close" lemma="the" pos="D">
   <span> <target id="w286"/> </span>
  </term>
  <term tid="t205" type="open" lemma="Sub-Group" pos="O">
   <span> <target id="w287"/> </span>
  </term>
  <term tid="t206" type="open" lemma="chairman" pos="N">
   <span> <target id="w288"/> </span>
  </term>
  <term tid="t207" type="open" lemma="and" pos="O">
   <span> <target id="w289"/> </span>
  </term>
  <term tid="t208" type="close" lemma="their" pos="D">
   <span> <target id="w290"/> </span>
  </term>

…

 <term tid="t227" type="open" lemma="them" pos="N">
   <span> <target id="w310"/> </span>
  </term>
  <term tid="t228" type="open" lemma="warmly" pos="A">
   <span> <target id="w311"/> </span>
  </term>
  <term tid="t229" type="close" lemma="for" pos="P">
   <span> <target id="w312"/> </span>
  </term>
  <term tid="t230" type="close" lemma="their" pos="D">
   <span> <target id="w313"/> </span>
  </term>
  <term tid="t231" type="open" lemma="effort" pos="N">
   <span> <target id="w314"/> </span>
  </term>
</terms>
<deps>
  <dep from="t206" to="t205" rfunc="mod"/>

…

  <dep from="t226" to="t227" rfunc="dobj"/>
</deps>
<chunks>
  <chunk cid="c171" head="t204" phrase="D">
      <span> <target id="t204"/> </span>
  </chunk>
  <chunk cid="c173" head="t206" phrase="NP">  <!—chairman-->
      <span> <target id="t206"/> </span>
  </chunk>

…

  <chunk cid="c191" head="t227" phrase="NP">  <!—them -->
      <span> <target id="t227"/> </span>
  </chunk>
  <chunk cid="c192" head="t228" phrase="A">
      <span> <target id="t228"/> </span>
  </chunk>
  <chunk cid="c194" head="t230" phrase="D">
      <span> <target id="t230"/> </span>
  </chunk>
  <chunk cid="c195" head="t231" phrase="NP">  <!—effort -->
      <span> <target id="t231"/> </span>
  </chunk>
</chunks>
</KAF>
```

Figure 5.1. *Words Extractor's input KAF file*

Figure 5.2. corresponds to the obvious output of Words Extractor, e.g. the original sentence.

```
The Sub-Group chairmen and their staff have gone about their work with determination, enthusiasm
and energy and I take this opportunity to thank them warmly for their efforts.
```

Figure 5.2. *Words Extractor's output Words_File1*

5.2.2.- *BART execution*

A preliminary work must be done before we can execute BART. Downloading, installing and configuring the service allowed us to use BART in an user friendly web based way or in console mode. (Note: Steps to install, configure and run BART in Appendix.)

Although BART is primarily meant as a platform for experimentation, it can be used simply as a coreference resolver. Since it is possible to import raw text, perform preprocessing and coreference resolution, and finally export the results to in-line XML format, we decided to exploit this characteristic of the tool and so, use it just like it was.

Figure 5.3. shows BART's output with the coreference relations proposed for the sentence.

```
<text>
 <s>
  <coref set-id="set_0">
    <w pos="dt">The</w>
    <w pos="jj">Sub-Group</w>
    <w pos="nns">chairmen</w>
  </coref>
  <w pos="cc">and</w>
  <coref set-id="set_0">
    <w pos="prp$">their</w>
  </coref>
  <w pos="nn">staff</w>
  <w pos="vbp">have</w>
  <w pos="vbn">gone</w>
  <w pos="in">about</w>
  <coref set-id="set_0">
     <w pos="prp$">their</w>
  </coref>
  <w pos="nn">work</w>
  <w pos="in">with</w>
  <w pos="nn">determination</w>
  <w pos=",">,</w>
  <w pos="nn">enthusiasm</w>
  <w pos="cc">and</w>
  <w pos="nn">energy</w>
  <w pos="cc">and</w>
  <w pos="prp">I</w>
  <w pos="vb">take</w>
  <w pos="dt">this</w>
  <w pos="nn">opportunity</w>
  <w pos="to">to</w>
  <w pos="vb">thank</w>
  <coref set-id="set_0">
    <w pos="prp">them</w>
  </coref>
  <w pos="rb">warmly</w>
  <w pos="in">for</w>
  <coref set-id="set_0">
    <w pos="prp$">their</w>
  </coref>
  <w pos="nns">efforts</w>
```

```
  <w pos=".">.</w>
 </s>
</text>
```

Figure 5.3. BART's output Words_File2 (includes coreferences)

5.2.3.- Treating token mismatches

Before any further processing can be done, a text needs to be segmented into words and sentences. This process is called tokenization. Tokenization is the initial phase in any NLP work, that is, it is the process of breaking up a given text into units called tokens. This tokens may be words, numbers, punctuation marks, parentheses or quotation marks and they are the basic units which will be decomposed in subsequent processing.

There are different techniques to accomplish the task of tokenization and the characteristics of a language are determinant to choose one method of tokenization or another. The fundamental difference in tokenization techniques lies on the fact of being an alphabetic language like English or an ideographic language like Chinese. The tokenization of alphabetic and ideographic languages are actually two rather different tasks which require different methods. While Ideographic languages provide no comparable information about sentence boundaries (which makes tokenization a much harder task), alphabetic languages usually separate words by blanks, and a tokenizer which simply replaces whitespace with word boundaries and cuts off punctuation marks, parentheses, and quotation marks at both ends of a word, is already quite accurate. Nevertheless, there still are some problems which need to be solved like punctuation disambiguation, locale expressions, multiword expressions, clitics or dehyphenation, just to mention some of them. Is in the way of dealing with these problems that we have found some differences between the tokenization process in Kyoto and in BART and for this reason, a work to assure that the two tokenizers match up with the tokens proposed inside the coreference have been done.

In the example we are following, no differences have been found between the tokens given by the two tokenizers. Otherwise, our system would reject the proposed coreference and go on to treat the next coreference. So, the output of this task is exactly the same as its input.

Figure 5.4. shows the accepted coreferences output *CorefsWords_File* after contrasting that there are no tokenization differences for this sentence between BART and Kyoto.

```
<text>
 <s>
  <coref set-id="set_0">
    <w pos="dt">The</w>
    <w pos="jj">Sub-Group</w>
    <w pos="nns">chairmen</w>
  </coref>
  <w pos="cc">and</w>
  <coref set-id="set_0">
    <w pos="prp$">their</w>
  </coref>
  <w pos="nn">staff</w>
  <w pos="vbp">have</w>
  <w pos="vbn">gone</w>
```

```
  <w pos="in">about</w>
  <coref set-id="set_0">
    <w pos="prp$">their</w>
  </coref>
  <w pos="nn">work</w>
  <w pos="in">with</w>
  <w pos="nn">determination</w>
  <w pos=",">,</w>
  <w pos="nn">enthusiasm</w>
  <w pos="cc">and</w>
  <w pos="prp">I</w>
  <w pos="vb">take</w>
  <w pos="dt">this</w>
  <w pos="nn">opportunity</w>
  <w pos="to">to</w>
  <w pos="vb">thank</w>
  <coref set-id="set_0">
    <w pos="prp">them</w>
  </coref>
  <w pos="rb">warmly</w>
  <w pos="in">for</w>
  <coref set-id="set_0">
    <w pos="prp$">their</w>
  </coref>
  <w pos="nns">efforts</w>
  <w pos=".">.</w>
 </s>
</text>
```

Figure 5.4. Output CorefsWords_File with the accepted coreferences

5.2.4.- Terms conversion

After rejecting the coreferences that didn't match with KAF tokens, the remaining coreferences (which are expressed as inflected forms) must be converted to their canonical forms (called 'terms' in KAF) for Kyoto to take advantage of them. This is quite a simple process because all the work that must be done consists in controlling the word's position in the original text and look for its identification number in the KAF file with respect to that position. Once the word's id number is obtained, the only work left to do will be extracting the term and the information related to that term, basically Part Of Speech Information.

Figure 5.5. shows the accepted coreferences "*The Sub-Group Chairmen*", "*their*", "*their*", "*them*" and "their".

```
  <coref set-id="set_0">
    <w pos="dt">The</w>
    <w pos="jj">Sub-Group</w>
    <w pos="nns">chairmen</w>
  </coref>

  <coref set-id="set_0">
    <w pos="prp$">their</w>
  </coref>

  <coref set-id="set_0">
    <w pos="prp$">their</w>
  </coref>

  <coref set-id="set_0">
    <w pos="prp">them</w>
  </coref>
```

```
  <coref set-id="set_0">
    <w pos="prp$">their</w>
  </coref>
```

Figure 5.5. coreferences of CorefsWords_File

The corresponding word identification numbers for these coreferences in KAF are *w286, w287, w288, w290, w295, w310 and w313*, and the terms related to these words are *"the sub-group chairman", "their", their", "them"*, and *"their"*, respectively.

Figure 5.6 gives a snapshot of the words as mentioned in KAF (refer to Figure 5.2. to see the KAF file).

```
…

  <wf wid="w286" sent="15" page="1">The</wf>
  <wf wid="w287" sent="15" page="1">Sub-Group</wf>
  <wf wid="w288" sent="15" page="1">chairmen</wf>

…

  <wf wid="w290" sent="15" page="1">their</wf>

…

  <wf wid="w295" sent="15" page="1">their</wf>

…

  <wf wid="w310" sent="15" page="1">them</wf>

…

  <wf wid="w313" sent="15" page="1">their</wf>

…
```

Figure 5.6. KAF file extract with the words identification numbers

Figure 5.7 gives a snapshot of the terms as mentioned in KAF (refer to Figure 5.2. to see the KAF file).

```
…

 <term tid="t204" type="close" lemma="the" pos="D">
   <span> <target id="w286"/> </span>
  </term>
  <term tid="t205" type="open" lemma="Sub-Group" pos="O">
   <span> <target id="w287"/> </span>
  </term>
  <term tid="t206" type="open" lemma="chairman" pos="N">
   <span> <target id="w288"/> </span>
  </term>

…

  <term tid="t208" type="close" lemma="their" pos="D">
   <span> <target id="w290"/> </span>
  </term>

…

  <term tid="t213" type="close" lemma="their" pos="D">
```

```
   <span> <target id="w295"/> </span>
  </term>

…

  <term tid="t227" type="open" lemma="them" pos="N">
   <span> <target id="w310"/> </span>
  </term>

…

  <term tid="t230" type="close" lemma="their" pos="D">
   <span> <target id="w313"/> </span>
  </term>

…
```

Figure 5.7. KAF file extract with the terms related to the words identification numbers

Figure 5.8. shows the coreferences list with the word `chairmen` substituted by its canonical form `chairman`.

```
  <coref set-id="set_0">
    <w pos="dt">The</w>
    <w pos="jj">Sub-Group</w>
    <w pos="nns">chairman</w>
  </coref>

…

  <coref set-id="set_0">
    <w pos="prp$">their</w>
  </coref>

…

  <coref set-id="set_0">
    <w pos="prp$">their</w>
  </coref>

…

  <coref set-id="set_0">
    <w pos="prp">them</w>
  </coref>

…

  <coref set-id="set_0">
    <w pos="prp$">their</w>
  </coref>
```

Figure 5.8. coreferences of *CorefsTerms_File* where the word `chairmen` has been substituted by the term `chairman`.

5.2.5.- Converting coreference chains into antecedent-anaphor pairs

Coreference chain is the set of coreferent referring expressions in a discourse. As we have shown in previous figures, BART offers us chains of expressions that are coreferent between them. In contrast, anaphora is the coreference of one referring expression with its antecedent. This referring expression is called anaphora and often it is a personal, a possessive or a demonstrative pronoun which refers back to something mentioned previously, called antecedent.

As a first approach to incorporate coreference resolution in Kyoto project, we decided to manage only the coreferences of BART that had pronouns between their referring expressions and also, we decided to organize and integrate these coreferences in the KAF file as pronominal anaphora. Usually, both the antecedent and the anaphor are used as referring expressions and having the same referent in the real world, they are termed coreferential.

For this purpose, we developed a subtask that gathered the coreferent expressions, selected the pronoun type referring expressions and, got for each of them the first antecedent (namely the first noun phrase) that was found previous to its appearance in the sentence. This way, our system constructed pronominal anaphoric pairs newly organized.

Figure 5.9. shows delimited by different colors the four pairs of antecedent-anaphor proposed by our system based on the initially given coreference chain of our example. As it can be seen, the first element of the pair corresponds to a noun phrase (antecedent) while the second element corresponds to a pronominal element (anaphor).

```
<coref set-id="set_0">
  <w pos="dt">The</w>
  <w pos="jj">Sub-Group</w>
  <w pos="nns">chairman</w>
</coref>

<coref set-id="set_0">
  <w pos="prp$">their</w>
</coref>

<coref set-id="set_0">
  <w pos="dt">The</w>
  <w pos="jj">Sub-Group</w>
  <w pos="nns">chairman</w>
</coref>

<coref set-id="set_0">
  <w pos="prp$">their</w>
</coref>

<coref set-id="set_0">
  <w pos="dt">The</w>
  <w pos="jj">Sub-Group</w>
  <w pos="nns">chairman</w>
</coref>

<coref set-id="set_0">
  <w pos="prp">them</w>
</coref>

<coref set-id="set_0">
  <w pos="dt">The</w>
  <w pos="jj">Sub-Group</w>
  <w pos="nns">chairman</w>
</coref>

<coref set-id="set_0">
  <w pos="prp$">their</w>
</coref>
```

Figure 5.9. Our coreference chain converted into antecedent-anaphor pairs

5.2.6.- Inserting pronominal anaphora information in the original KAF File

As we previously said, a KAF file is a XML annotation format that incorporates standardized proposals for the linguistic annotation of text but represents them in an easy to use layered structure.
Once our system has converted the coreference chains into pronominal anaphoric information, the only work left to do will be to give identification numbers to each antecedent-anaphoric pair and to put the pairs into the KAF file along with the words, terms, constituents and syntactic dependencies that have been annotated for the text and are stored separately, respecting the references across the structures.

Figure 5.10. shows the final KAF file. It now includes pronominal anaphoric information as part of its linguistic annotated information.

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<KAF xml:lang="en" doc="KyotoKAFEng_3/1286">
<text>
  <wf wid="w286" sent="15" page="1">The</wf>

…

  <wf wid="w315" sent="15" page="1">.</wf>
</text>
<terms>
 <term tid="t204" type="close" lemma="the" pos="D">
   <span> <target id="w286"/> </span>
  </term>

…

  <term tid="t231" type="open" lemma="effort" pos="N">
   <span> <target id="w314"/> </span>
  </term>
</terms>
<deps>
  <dep from="t206" to="t205" rfunc="mod"/>

…

 <dep from="t226" to="t227" rfunc="dobj"/>
</deps>
<chunks>
  <chunk cid="c171" head="t204" phrase="D">
      <span> <target id="t204"/> </span>
  </chunk>

…

  <chunk cid="c195" head="t231" phrase="NP">
      <span> <target id="t231"/> </span>
  </chunk>
</chunks>
<corefs>
  <coref set_id="0-1">
   <span>
    <target id = "t204" term = "the " pos_bart = "dt" pos_KAF = "D"/>
    <target id = "t205" term = "Sub-Group " pos_bart = "jj" pos_KAF = "O"/>
    <target id = "t206" term = "chairman " pos_bart = "nns" pos_KAF = "N"/>
  </span>
  <span>
    <target id = "t208" term = "their " pos_bart = "prp$" pos_KAF = "D"/>
  </span>
 </coref>
```

```
 <coref set_id="0-2">
  <span>
    <target id = "t204" term = "the " pos_bart = "dt" pos_KAF = "D"/>
    <target id = "t205" term = "Sub-Group " pos_bart = "jj" pos_KAF = "O"/>
    <target id = "t206" term = "chairman " pos_bart = "nns" pos_KAF = "N"/>
    </span>
  <span>
    <target id = "t213" term = "their " pos_bart = "prp$" pos_KAF = "D"/>
  </span>
 </coref>
 <coref set_id="0-3">
   <span>
    <target id = "t204" term = "the " pos_bart = "dt" pos_KAF = "D"/>
    <target id = "t205" term = "Sub-Group " pos_bart = "jj" pos_KAF = "O"/>
    <target id = "t206" term = "chairman " pos_bart = "nns" pos_KAF = "N"/>
  </span>
  <span>
    <target id = "t227" term = "them " pos_bart = "prp" pos_KAF = "N"/>
  </span>
 </coref>
 <coref set_id="0-4">
   <span>
     <target id = "t204" term = "the " pos_bart = "dt" pos_KAF = "D"/>
     <target id = "t205" term = "Sub-Group " pos_bart = "jj" pos_KAF = "O"/>
     <target id = "t206" term = "chairman " pos_bart = "nns" pos_KAF = "N"/>
   </span>
   <span>
     <target id = "t230" term = "their " pos_bart = "prp$" pos_KAF = "D"/>
   </span>
  </coref>
</corefs>
</KAF>
```

Figure 5.10. Final output. The newly created KAF file with pronominal anaphora information

6.- Evaluation

The effectiveness of the final result of this work depends mainly on an accurate coreference resolution and on the patterns constructed for the kybots. For this reason, the evaluation of this work is directly related to the good response of BART and also to the precise construction of patterns that will identify concept instances taking into account the semantic information given by our system.

We chose from KAF files a battery of sentences that had coreferent elements. Specifically, we chose sentences that had noun phrases and pronouns that referred to the same entity. The pronouns picked out were `their, they` and `it`. We used the sentences as input for BART and we evaluated the resulting output to measure the precision of the selected coreference resolver.
The following figures show classified by pronoun types the set of sentences gathered for this experiment. We show first for each pronoun type the sentences that were correctly analyzed by BART. The correct coreferent elements sugested in each sentence are underlined with a straight line. Next, we show the sentences that were incorrectly analyzed by BART. The incorrect coreferent elements sugested by BART in each sentence are underlined with a waved line. Also, for the incorrect sentences we mention between parentheses which one(s) should be the correct coreferent elements.

Figure 6.1. shows the analysis of BART for 'they' type sentences.

'THEY' SENTENCES:

Correct coreference resolutions:

√ *Messages and proposals from action are most likely to be received sympathetically if they come from leading and respected figures from the sector concerned.*

√ *This in turn requires a better understanding of ecosystems and how robust they are in responding to land use change and other impacts and perturbations.*

√ *We should conserve species and habitats because they enrich our lives.*

√ *Recently planted woodlands are less diverse, immature ecosystems, although they can add to the biodiversity of a previously unwooded environment, especially land of low wildlife value.*

√ *A community of interdependent organisms and the environment they inhabit, such as ponds and pond life.*

√ *Where local records centres exist they should be the centre of the network.*

√ *Examples include the use of water resources, the generation and use of energy, and transport systems, particularly roads and the traffic they carry.*

√ *For example, communication between modern databases can be achieved by specifying that they all use, or at least can be accessed through Structured Query Language (SQL).*

Incorrect coreference resolutions:

X *Survey methods are simple and efficient, and volunteers record details of both the birds[1] they[1] encounter and the habitats[2] they[2] live in. (volunteers, they[1] / the birds, they[2])*

X *In other cases, perhaps in the more commercially orientated public bodies and the private sector, the key datasets are almost by-products of their main activities and they need relatively little reciprocity from other custodians. (public bodies, their / datasets, they)*

Figure 6.1. Set of sentences extracted from KAF files with 'they' pronominal coreferent elements.

Figure 6.2. shows the analysis of BART for 'their' type sentences.

'THEIR' SENTENCES':

Correct coreference resolutions:

√ *The Sub-Group chairmen and their staff have gone about their work with determination, enthusiasm and energy and I take this opportunity to thank them warmly for their efforts.*

√ *The world biodiversity has been taken by many to mean not simply the variety of life forms on earth, but also the urgent need to ensure their survival.*

√ *For more information on the three lists, their contents, and the action plans see Annex F and Annex G.*

√ *Alpine rivers and the herbaceous vegetation along their banks.*

√ *The production of such strategies is voluntary and their status advisory.*

√ *Further work is required for a better understanding of these factors and their effects upon birds.*

√ *The key organisations concerned with this area are the statutory nature conservation agencies and their Joint Committee together with ITE, and in particular the Biological Records Centre, which is part-funded by the JNCC.*

Incorrect coreference resolutions:

X *As well as variation[1] between species[2], there is also variation[1] within species[2], reflecting their genetic make-up. (species, their)*

X *Natural selection and random changes in genetic composition in isolated patches of species will drive dynamic changes in their genetic structure resulting in distinctive local populations. (species, their)*

X *Local Biodiversity Action Plans provide a focus for local initiatives, and provide an opportunity for local people to express their views on what is important. (people, their)*

X *The UK is fortunate in still possessing a significant number of individuals who contribute vital data as a result of their enthusiasm for nature. (individuals, their)*

X *An important contribution can be made at both national and local levels by maximising the use of the Recorder sofware system for recording species and their geographical locations. (species, their)*

X *Custodians of information have a range of tools to assist access to their most used datasets. (Custodians, their)*

X *A crude summary is that recorders of data and organisations collating datasets acquire Intellectual Property Rights which may pass to their estate. (recorders, their)*

Figure 6.2. Set of sentences extracted from KAF files with 'their' pronominal coreferent elements.

Figure 6.3. shows the analysis of BART for 'it' type sentences.

'IT' SENTENCES:

Correct coreference resolutions:

√ *These will show major indicators relevant to wildlife and biodiversity and to the key impacts on it from man.*

Incorrect coreference resolutions:

X Avoiding duplication of effort and errors in copying data so that, ideally, each dataset is managed to agreed standards by a single known individual or organisation and made available to others who have a legitimate use for it. (dataset, it)

X But there are important gaps which it will be slow and expensive to fill. (gaps, it)

X The value of the system to individuals and small organisations would be increased if it could be loaded with appropriate national information, such as species of conservation concern or action plans, and could therefore provide both feedback and context for local work. (system, it) *NOTE:* BART does not find any coreferent elements.

X We consider that a UK Biodiversity Database (UKDB) would be an important tool for carrying forward the Biodiversity Action Plan and other commitments under UD and EC legislation and international conventions. At relatively little cost, it would add considerable value to the high volume of existing and planned data. (Biodiversity Action Plan, it) *NOTE:* BART does not find any coreferent elements.

X While a wide range of organisations produce environmental or conservation material, this information is not always practical enough nor does it always reach those practitioners who need it. (this information, it, it)

Figure 6.3. Set of sentences extracted from KAF files with 'it' pronominal coreferent elements.

Table 6.1 shows the overall results for the set of sentences analyzed for each pronoun.

Evaluated pronominals	Total evaluated sentences	Correct coreference resolution	Incorrect coreference resolution	Precision (%)
THEY	8	6	2	**75**
THEIR	14	7	7	**50**
IT	6	1	5	**16,66**
Total	*28*	*14*	*14*	*50*

Table 6.1. Evaluation results of BART

7- Conclusions and further work

Whereas the last 10 years have seen considerable advances in the field of coreference resolution, there are still a number of outstanding issues that remain either unsolved or need more attention and, as a consequence, represent major challenges to the further development of the field. Mayor research into the factors influencing the performance of the resolution algorithm is necessary.

Despite of being aware of the small size of the sample evaluated in this work, we think that we can conclude by looking at the results that BART responds better to the sentences that have coreferent elements with the pronouns 'their' or 'they' rather than to those sentences that have

coreferent elements that include pronouns of the 3rd person of the singular ('it' or 'its'). This in turn, is easy to understand because en english the treatment of coreferent elements that have pronouns like 'it' or 'its' is much more difficult to solve. For this reason, we think that we could take advantage of the fact that BART responds considerably well with pronouns like 'their' or 'they'. Therefore, we could use it into Kyoto project in general and into the kybots in particular to improve the information extraction system detecting more and better concept instances and relations in the text.

Nevertheless, it could be possible to adapt BART to other classifiers and features. It is usually possible to achieve the bulk of this task by simply mixing existing components for preprocessing and feature extraction, which can be made modifying only configuration settings and an XML based description of the feature set and learner(s) used.

That is why we think that, on one hand, future work lies on trying different configuration settings and combining other features, and learning components within BART. This way, new evaluations and comparisons of results could be done and so, could be determined the best conditions to execute this coreference resolver.

On the other hand, work to define specific patterns for kybots that use the semantic information appended to KAF files should be done. Only this way could be possible the identification of more and better concept instances and relations in the texts and more precise and accurate outcome could be achieved in general by the tools developed under Kyoto project.

8.- References

Banko, M., O. Etzioni, S. Soderland, D. and Weld. 2008. Open information extraction from the web.

Bean, D., and E. Riloff. 2004. Unsupervised Learning of Contextual Role Knowledge for Coreference Resolution.

Bengtson,E., and D. Roth. 2008. Understanding the value of features for coreference resolution.

Blanchard, e., and D. Allard. 2010. Handbook of Research on Culturally-Aware Information Technology: Perspectives and Models.

Bosma, W., P. Vossen, A. Soroa, G. Rigau, M. Tesconi, A. Marchetti, M. Monachini, and C. Aliprandi. 2009. KAF: a generic semantic annotation format.

Charniak, E., and M. Johnson. 2005. Coarse-to-fine n-best parsing and maxent discriminative reranking.

Christensen, J., Mausam, S. Soderland, and O. Etzioni. 2011. An Analysis of Open Information Extraction based on Semantic Role Labeling.

Cunningham, H., D. Maynard, K. Bontcheva, V. Tablan. 2002. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications.

Cunningham, H., D. Maynard, and V. Tablan. 2000. JAPE: a Java Annotation Patterns Engine.

Etzioni, O., A. Fader, J. Christensen, S. Soderland, and Mausam. 2011. Open information Extraction: the Second Generation.

Etzioni, O., M. Cafarella, D. Downey, S. Kok, A. Popescu, T. Shaked, S. soderland, D. Weld, and A. Yates. 2005. Unsupervied named-entity extraction from the web: An experimental study.

Fader, A., Soderland, S. and O. Etzioni, 2011. Identifying Relations for Open Information Extraction

Finkel, J., T. Grenager, and C. Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling.

Haghighi, A., D. Klein. 2010. An Entity-Level Approach to Information Extraction.

Haghighi, A., D. Klein. 2009. Simple Coreference Resolution with Rich Syntactic and Semantic Features.

Kabadjov, M. 2007. A Comprehensive Evaluation of Anaphora Resolution and Discourse-new Recognition.

Kobdani, H., and H. Schutze . 2010. SUCRE: A Modular System for coreference Resolution. SemEval.

Kudoh, T. and Y. Matsumoto. 2000. Use of Support Vector Machines for chunk identification.

Lee, H., Y. Perisman, A. Chang, Chambers, M. Surdeanu, and D. Jurafsky. 2011. Stanford's Multi-Pass Sieve coreference Resolution System. CoNLL.

Lei, Y., S. Uren, and E. Motta. 2006. SemSearch: A Search Engine for the Semantic Web.

Lopez, V., Pasin, M., and Motta, E. AquaLog: An Ontology-portable Question Answering System for the Semantic Web.

Moschitti, A. 2006. Making tree kernels practical for natural language learning.

Ng, V., and C. Cardie. 2002a. Improving machine learning approaches to coreference resolution.

Ng, V., and C. Cardie. 2002b. Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution.

Poesio, M., M. Kabadjov. 2004. A general-purpose, off the shelf anaphoric resolver.

Ponzetto, S., and M. Strube. 2006. Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution.

Popov, B., A. Kiryakov, D. Manov, D. Ognyanoff and M. Goranov. 2003. KIM – Semantic Annotation Platform.

Recasens, M. and M. Vila. 2000. On Paraphrase and coreference. Computational Linguistics Volume 36 Issue 4.

Shinyama, Y., and S. Sekine. 2006. Preemptive Information Extraction using Unrestricted Relation Discovery.

Soon et al. 2001. A Machine Learning Approach to coreference resolution of Noun Phrases. ACL.

Stoyanov et al. 2010. Coreference Resolution with Reconcile. ACL.

Tablan, V., T. Plajnar, H. Cunningham, and K. Bontcheva. 2006. User-friendly ontology authoring using a controlled language. LREC.

Toutanova, K., D. Klein, C. Manning, and Y. Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network.

Versley Y., S. Ponzetto, M. Poesio, V. Eidelman, A. Jern, J. Smith, X. Yang, and A. Moschitti. 2008. BART: A modular toolkit for coreference resolution.

Versley, Y. 2006. A constraint-based approach to noun phrase coreference resolution in German

newspaper text.

Vicient, C. 2011. Ontology-based information Extraction. Master Thesis.

Vincent, Ng. 2010. Supervised Noun Phrase Coreference Research: The First Fifteen Years. ACL.

Witten, I., and E. Frank. 2005. Data Mining: Practical machine learning tools and techniques.

Wu, F., and D. Weld. 2010. Open information extraction using Wikipedia.

Yang, X., J. Su, J., and C. Tan. 2006. Kernel-based pronoun resolution with structured syntactic knowledge.

Yang, X., and J. Su. 2007. Coreference Resolution Using Semantic Relatedness Information from Automatically Discovered Patterns. ACL.

Yang, X., J. Su, J. Lang , C. Lim , T. Ting , and L. Li. 2008. An Entity-Mention Model for coreference Resolution with Inductive Logic Programming. ACL.

Zhou, C., et al. 2006. Approaches to Text Mining for Clinical Medical Records.

Appendix

Steps to install and run BART tool:

- Download BART-snapshot.tgz tarball and untar it somewhere.

- Change to that directory and do:

  ```
  > source setup.sh
  > java -Xmx1024m elkfed.webdemo.BARTServer &
  ```

  (the first command sets up the classpath, whereas the second starts BART's web service).

- Point the browse at http://localhost:8125/index.jsp and enter some text into the from and verify that it does something. Clicking on the “coref” tab should run the coreference resolver and display markables that are part of a coreference chain.

- To use BART on larger quantities of text, you would want to use from the terminal this other command:

  ```
  > nohup cat Words_File1 | POST -t 100000
  http://localhost:8125/BARTDemo/ShowText/process/ > XMLoutputFile.txt &
  ```

Figure 5.1.

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<KAF xml:lang="en" doc="KyotoKAFEng_3/1286">
<text>
  <wf wid="w286" sent="15" page="1">The</wf>
  <wf wid="w287" sent="15" page="1">Sub-Group</wf>
  <wf wid="w288" sent="15" page="1">chairmen</wf>
  <wf wid="w289" sent="15" page="1">and</wf>
  <wf wid="w290" sent="15" page="1">their</wf>
  <wf wid="w291" sent="15" page="1">staff</wf>
  <wf wid="w292" sent="15" page="1">have</wf>
  <wf wid="w293" sent="15" page="1">gone</wf>
  <wf wid="w294" sent="15" page="1">about</wf>
  <wf wid="w295" sent="15" page="1">their</wf>
  <wf wid="w296" sent="15" page="1">work</wf>
  <wf wid="w297" sent="15" page="1">with</wf>
  <wf wid="w298" sent="15" page="1">determination</wf>
  <wf wid="w299" sent="15" page="1">,</wf>
  <wf wid="w300" sent="15" page="1">enthusiasm</wf>
  <wf wid="w301" sent="15" page="1">and</wf>
  <wf wid="w302" sent="15" page="1">energy</wf>
  <wf wid="w303" sent="15" page="1">and</wf>
  <wf wid="w304" sent="15" page="1">I</wf>
  <wf wid="w305" sent="15" page="1">take</wf>
  <wf wid="w306" sent="15" page="1">this</wf>
  <wf wid="w307" sent="15" page="1">opportunity</wf>
  <wf wid="w308" sent="15" page="1">to</wf>
  <wf wid="w309" sent="15" page="1">thank</wf>
  <wf wid="w310" sent="15" page="1">them</wf>
  <wf wid="w311" sent="15" page="1">warmly</wf>
  <wf wid="w312" sent="15" page="1">for</wf>
  <wf wid="w313" sent="15" page="1">their</wf>
  <wf wid="w314" sent="15" page="1">efforts</wf>
  <wf wid="w315" sent="15" page="1">.</wf>
</text>
<terms>
 <term tid="t204" type="close" lemma="the" pos="D">
   <span>
    <target id="w286"/>
   </span>
  </term>
  <term tid="t205" type="open" lemma="Sub-Group" pos="O">
   <span>
    <target id="w287"/>
   </span>
  </term>
  <term tid="t206" type="open" lemma="chairman" pos="N">
   <span>
    <target id="w288"/>
   </span>
  </term>
  <term tid="t207" type="open" lemma="and" pos="O">
   <span>
    <target id="w289"/>
   </span>
```

```
 </term>
 <term tid="t208" type="close" lemma="their" pos="D">
  <span>
   <target id="w290"/>
  </span>
 </term>
 <term tid="t209" type="open" lemma="staff" pos="N">
  <span>
   <target id="w291"/>
  </span>
 </term>
 <term tid="t210" type="open" lemma="have" pos="V">
  <span>
   <target id="w292"/>
  </span>
 </term>
 <term tid="t211" type="open" lemma="go" pos="V">
  <span>
   <target id="w293"/>
  </span>
 </term>
 </term>
 <term tid="t212" type="close" lemma="about" pos="P">
  <span>
   <target id="w294"/>
  </span>
 </term>
 <term tid="t213" type="close" lemma="their" pos="D">
  <span>
   <target id="w295"/>
  </span>
 </term>
 <term tid="t214" type="open" lemma="work" pos="N">
  <span>
   <target id="w296"/>
  </span>
 </term>
 <term tid="t215" type="close" lemma="with" pos="P">
  <span>
   <target id="w297"/>
  </span>
 </term>
 <term tid="t216" type="open" lemma="determination" pos="N">
  <span>
   <target id="w298"/>
  </span>
 </term>
 <term tid="t217" type="open" lemma="enthusiasm" pos="N">
  <span>
   <target id="w300"/>
  </span>
 </term>
<term tid="t218" type="open" lemma="and" pos="O">
  <span>
   <target id="w301"/>
  </span>
```

```
 </term>
 <term tid="t219" type="open" lemma="energy" pos="N">
  <span>
   <target id="w302"/>
  </span>
 </term>
 <term tid="t220" type="open" lemma="and" pos="O">
  <span>
   <target id="w303"/>
  </span>
 </term>
 <term tid="t221" type="open" lemma="I" pos="N">
  <span>
   <target id="w304"/>
  </span>
 </term>
 <term tid="t222" type="open" lemma="take" pos="V">
  <span>
   <target id="w305"/>
  </span>
 </term>
 <term tid="t223" type="close" lemma="this" pos="D">
  <span>
   <target id="w306"/>
  </span>
 </term>
 <term tid="t224" type="open" lemma="opportunity" pos="N">
  <span>
   <target id="w307"/>
  </span>
 </term>
 <term tid="t225" type="close" lemma="to" pos="P">
  <span>
   <target id="w308"/>
  </span>
 </term>
 <term tid="t226" type="open" lemma="thank" pos="V">
  <span>
   <target id="w309"/>
  </span>
 </term>
<term tid="t227" type="open" lemma="them" pos="N">
  <span>
   <target id="w310"/>
  </span>
 </term>
 <term tid="t228" type="open" lemma="warmly" pos="A">
  <span>
   <target id="w311"/>
  </span>
 </term>
 <term tid="t229" type="close" lemma="for" pos="P">
  <span>
   <target id="w312"/>
  </span>
 </term>
```

```
  <term tid="t230" type="close" lemma="their" pos="D">
   <span>
    <target id="w313"/>
   </span>
  </term>
  <term tid="t231" type="open" lemma="effort" pos="N">
   <span>
    <target id="w314"/>
   </span>
  </term>
</terms>
<deps>
  <dep from="t206" to="t205" rfunc="mod"/>
  <dep from="t209" to="t210" rfunc="subj"/>
  <dep from="t217" to="t222" rfunc="subj"/>
  <dep from="t219" to="t222" rfunc="subj"/>
  <dep from="t221" to="t222" rfunc="subj"/>
  <dep from="t222" to="t224" rfunc="dobj"/>
  <dep from="t226" to="t227" rfunc="dobj"/>
</deps>
<chunks>
  <chunk cid="c171" head="t204" phrase="D">
      <span>
        <target id="t204"/>
      </span>
  </chunk>
  <chunk cid="c173" head="t206" phrase="NP">
      <span>
        <target id="t206"/>
      </span>
  </chunk>
  <chunk cid="c175" head="t208" phrase="D">
      <span>
        <target id="t208"/>
      </span>
  </chunk>
  <chunk cid="c176" head="t209" phrase="NP">
      <span>
        <target id="t209"/>
      </span>
  </chunk>
  <chunk cid="c177" head="t211" phrase="VP">
      <span>
        <target id="t210"/>
      </span>
      <span>
        <target id="t211"/>
      </span>
  </chunk>
  <chunk cid="c179" head="t213" phrase="D">
      <span>
        <target id="t213"/>
      </span>
  </chunk>
  <chunk cid="c180" head="t214" phrase="NP">
      <span>
```

```
      <target id="t214"/>
    </span>
    <span>
      <target id="t215"/>
    </span>
    <span>
      <target id="t216"/>
    </span>
</chunk>
<chunk cid="c181" head="t217" phrase="NP">
    <span>
      <target id="t217"/>
    </span>
</chunk>
<chunk cid="c183" head="t219" phrase="NP">
    <span>
      <target id="t219"/>
    </span>
</chunk>
<chunk cid="c185" head="t221" phrase="NP">
    <span>
      <target id="t221"/>
    </span>
</chunk>
<chunk cid="c186" head="t222" phrase="VP">
    <span>
      <target id="t222"/>
    </span>
</chunk>
<chunk cid="c187" head="t223" phrase="D">
    <span>
      <target id="t223"/>
    </span>
</chunk>
<chunk cid="c188" head="t224" phrase="NP">
    <span>
      <target id="t224"/>
    </span>
  </chunk>
<chunk cid="c190" head="t226" phrase="VP">
    <span>
      <target id="t226"/>
    </span>
</chunk>
<chunk cid="c191" head="t227" phrase="NP">
    <span>
      <target id="t227"/>
    </span>
</chunk>
<chunk cid="c192" head="t228" phrase="A">
    <span>
      <target id="t228"/>
    </span>
</chunk>
<chunk cid="c194" head="t230" phrase="D">
    <span>
```

```
        <target id="t230"/>
      </span>
  </chunk>
  <chunk cid="c195" head="t231" phrase="NP">
      <span>
        <target id="t231"/>
      </span>
  </chunk>
</chunks>
</KAF>
```