Another Evaluation of Anaphora Resolution Algorithms and a Comparison with GETARUNS' Knowledge Rich Approach

Rodolfo Delmonte, Antonella Bristot, Marco
Aldo Piccolino Boniforti, Sara Tonelli
Department of Language Sciences
Università Ca' Foscari – Ca' Bembo
30120, Venezia, Italy
delmont@unive.it

Abstract

In this paper we will present an evaluation of current state-of-the-art algorithms for Anaphora Resolution based on a segment of Susanne corpus (itself a portion of Brown Corpus), a much more comparable text type to what is usually required at an international level for such application domains as Question/Answering, Information Extraction, Text Understanding, Language Learning. The portion of text chosen has an adequate size which lends itself to significant statistical measurements: it is portion A, counting 35,000 tokens and some 1000 third person pronominal expressions. The algorithms will then be compared to our system, GETARUNS, which incorporates an AR algorithm at the end of a pipeline of interconnected modules that instantiate standard architectures for NLP. F-measure values reached by our system are significantly higher (75%) than the other ones.

1 Introduction

The problem of anaphora resolution (hence AR) looms more and more as a prominent one in unrestricted text processing due to the need to recover semantically consistent information in most current NLP applications. This problem does not lend itself easily to a statistical approach so that rule-based approaches seem the only viable solution.

We present a new evaluation of three state-of-the-art algorithms for anaphora resolution – GuiTAR, JavaRAP, MARS – on the basis of a portion of Susan Corpus (derived from Brown Corpus) a much richer testbed than the ones previously used for evaluation, and in any case a much more comparable source with such texts as newspaper articles and stories. Texts used previously ranged from scientific manuals to descriptive scientific texts and were generally poor on pronouns and rich on nominal descriptions. Two of the algorithms – GuiTAR and JavaRAP - use Charniak's parser output, which contributes to the homogeneity of the type of knowledge passed to the resolution procedure. MARS, on the contrary, uses a more sophisticated input, the one provided by Connexor FDG-parser. The algorithms will then be compared to our system, GETARUNS, which incorporated an AR algorithm at the end of a pipeline of interconnected modules that instantiate standard architectures for NLP. The version of the algorithm presented here is a newly elaborated one, and is devoted to unrestricted text processing. It is an upgraded version from the one discussed in Delmonte (1999;2002a;2002b) and tries to incorporate as much as possible of the more sophisticated version implemented in the complete GETARUN (see Delmonte 1990;1991;1992;1994; 2003;2004).

The paper is organized as follows: in section 2 below we briefly discuss architectures and criteria for AR of the three algorithms evaluated. In section 3 we present our system. Section 4 is dedicated to a compared evaluation and a general discussion.

2 The Anaphora Resolution Algorithms

We start by presenting a brief overview of three state-of-the-art algorithms for anaphora resolution – GuiTAR, JavaRAP, MARS.

2.1 JavaRAP

As reported by the authors (Long Qiu, Min-Yen Kan, Tat-Seng Chua, 2004) of the JAVA implementation, head-dependent relations required by RAP are provided by looking into the structural "argument domain" for arguments and into the structural "adjunct domain" for adjuncts. Domain information is important to establish disjunction relations, i.e. to tell whether a third person pronoun can look for antecedents within a certain structural domain or not. According to Binding Principles, Anaphors (i.e. reciprocal and reflexive pronouns), must be bound – search for their binder-antecedent – in their same binding domain – roughly corresponding to the notion of structural "argument/adjunct domain". Within the same domains, Pronouns must be free. Head-argument or head-adjunct relation is determined whenever two or more NPs are sibling of the same VP.

Additional information is related to agreement features, which in the case of pronominal expressions are directly derived. As for nominal expressions, features are expressed in case they are either available on the verb – for SUBJect NPs– or else if they are expressed on the noun and some other tricks are performed for conjoined nouns. Gender is looked up in the list of names available on the web. This list is also used to provide the semantic feature of animacy.

RAP is also used to find pleonastic pronouns, i.e. pronouns which have no referents. To detect conditions for pleonastic pronouns a list of patterns is indicated, which used both lexical and structural information.

Salience weight is produced for each candidate antecedent from a set of salience factors. These factors include main Grammatical Relations, Headedness, non Adverbiality, belonging to the same sentence. The information is computed again by RAP, directly on the syntactic structure. The weight computed for each noun phrase is divided by two in case the distance from the current sentence increases. Only NPs contained within a distance of three sentences preceding the anaphor are considered by JavaRAP.

2.2 GuiTAR

The authors (Poesio, M. and Mijail A. Kabadjov 2004) present their algorithm as an attempt at providing a domain independent anaphora resolution module, "that developers of NLE applications can pick off the shelf in the way of tokenizers, POS taggers, parsers, or Named Entity classifiers". For these reasons, GuiTAR has been designed to be as independent as possible from other modules, and to be as modular as possible, thus "allowing for the possibility of replacing specific components (e.g., the pronoun resolution component)".

The authors have also made an attempt at specifying what they call the Minimal Anaphoric Syntax (MAS) and have devised a markup language based on GNOME mark-up scheme. In MAS, Nominal Expressions constitute the main processing units, and are identified with the tag NE <ne>, which have a CAT attribute, specifying the NP type: the-np, pronoun etc., as well as Person, Number and Gender attributes for agreement features. Also the internal structure of the NP is marked with Mod and NPHead tags.

The pre-processing phase uses a syntactic guesser which is a chunker of NPs based on heuristics. All NEs add up to a discourse model – or better History List - which is then used as the basic domain where Discourse Segments are contained. Each Discourse Segment in turn may be constituted by one or more Utterances. Each Utterance in turn contains a list of forward looking centers Cfs.

The Anaphora Resolution algorithm implemented is the one proposed by MARS which will be commented below. The authors also implemented a simple algorithm for resolving Definite Descriptions on the basis of the History List by a same head matching approach.

2.3 MARS

The approach is presented as a knowledge poor anaphora resolution algorithm (Mitkov R. [1995;1998]), which makes use of POS and NP chunking, it tries to individuate pleonastic "it" occurrences, and assigns animacy. The weighting algorithm seems to contain the most original approach. It is organized with a filtering approach by a series of indicators that are used to boost or reduce the score for antecedenthood to a given NP. The indicators are the following ones:

FNP (First NP); INDEF (Indefinite NP); IV (Indicating Verbs); REI (Lexical Reiteration); SH (Section Heading Preference); CM (Collocation Match); PNP (Prepositional Noun Phrases); IR (Immediate Reference); SI (Sequential Instructions); RD (Referential Distance); TP (Term Preference),

As the author comments, antecedent indicators (preferences) play a decisive role in tracking down the antecedent from a set of possible candidates. Candidates are assigned a score (-1, 0, 1 or 2) for each indicator; the candidate with the highest aggregate score is proposed as the antecedent.

The authors comment is that antecedent indicators have been identified empirically and are related to salience (definiteness, givenness, indicating verbs, lexical reiteration, section heading preference, "non- prepositional" noun phrases), to structural matches (collocation, immediate reference), to referential distance or to preference of terms. However it is clear that most of the indicators have been suggested for lack of better information, in particular no syntactic constituency was available.

In a more recent paper (Mitkov et al., 2003) MARS has been fully reimplemented and the indicators updated. The authors seem to acknowledge the fact that anaphora resolution is a much more difficult task than previous work had suggested, In unrestricted text analysis, the tasks involved in the anaphora resolution process contribute a lot of uncertainty and errors that may be the cause for low performance measures.

The actual algorithm uses the output of Connexor's FDG Parser, filters instances of "it" and eliminates pleonastic cases, then produces a list of potential antecedents by extracting nominal and pronominal heads from NPs preceding the pronoun. Constraints are then applied to this list in order to produce the "set of competing candidates" to be considered further, i.e. those candidates that agree in number and gender with the pronoun, and also obey syntactic constraints. They also introduced the use of Genetic Algorithms in the evaluation phase.

The new version of MARS includes three new indicators which seem more general and applicable to any text, so we shall comment on them.

Frequent Candidates (FC) – this is a boosting score for most frequent three NPs; Syntactic Parallelism (SP) – this is a boosting score for NPs with the same syntactic role as the pronoun, roles provided by the FDG-Parser; Boost Pronoun (BP) – pronoun candidates are given a bonus (no indication of conditions for such a bonus).

The authors also reimplemented in a significant way the indicator First NPs which has been renamed, "Obliqueness (OBL) – score grammatical functions, SUBJect > OBJect > IndirectOBJect > Undefined".

MARS has a procedure for automatically identifying pleonastic pronouns: the classification is done by means of 35 features organized into 6 types and are expressed by a mixture of lexical and grammatical heuristics. The output should be a fine-grained characterization of the phenomenon of the use of pleonastic pronouns which includes, among others, discourse anaphora, clause level anaphora and idiomatic cases.

In the same paper, the authors deal with two more important topics: syntactic constraints and animacy identification.

3 GETARUNS

In a number of papers (Delmonte 1990;1991; 1992;1994; 2003;2004) and in a book (Delmonte 1992) we described our algorithms and the theoretical background which inspired it. Whereas the old version of the system had a limited vocabulary and was intended to work only in limited domains with high precision, the current version of the system has been created to cope with unrestricted text. In Delmonte (2002), we reported preliminary results obtained on a corpus of anaphorically annotated texts made available by R.Mitkov on his website. Both definite descriptions and pronominal expressions were considered, success rate was at 75% F-measure. In those case we used a very shallow and robust parser which produced only NP chunks which were then used to fire anaphoric processes. However the texts making up the corpus were technical manuals, where the scope and usage of pronominal expressions is very limited.

The current algorithm for anaphora resolution works on the output of a complete deep robust parser which builds an indexed linear list of dependency structures where clause boundaries are clearly indicated; differently from Connexor, our system elaborates both grammatical relations and semantic roles information for arguments and adjuncts. Semantic roles are very important in the weighting procedures. Our system also produces implicit grammatical relations which are either controlled SUBJects of untensed clauses, arguments or adjuncts of relative clauses.

As to the anaphoric resolution algorithm, it is based on the original Sidner's (1983:Chapter 5) and Webber's (1983:Chapter 6) intuitions on Focussing in Discourse. We find distributed, local approaches to anaphora resolution more efficient than monolithic, global ones. In particular we believe that due to the relevance of structural constraints in the treatment of locally restricted classes of pronominal expressions, it is more appropriate to activate different procedures which by dealing separately with non-locally restricted classes also afford separate evaluation procedures. There are also at least two principled reasons for the separation into two classes.

The first reason is a theoretical one. Linguistic theory has long since established without any doubt the existence in most languages of the world of at least two classes: the class of pronouns which must be bound locally in a given domain and the class of pronouns which must be left free in the same domain – as a matter of fact, English also has a third class of pronominals, the so-called long-distance subject-of-consciousness bound pronouns (see Zribi-Hertz A., 1989);

The second reason is empirical. Anaphora resolution is usually carried out by searching antecedents backward w.r.t. the position of the current anaphoric expression. In our approach, we proceed in a clause by clause fashion, weighting each candidate antecedent w.r.t. that domain, trying to resolve it locally. Weighting criteria are amenable on the one hand to linear precedence constraints, with scores assigned on a functional/semantic basis. On the other hand, these criteria may be overrun by a functional ranking of clauses which requires to treat main clauses differently from secondary clauses, and these two differently from complement clauses. On the contrary, global algorithms neglect altogether such requirements: they weight each referring expression w.r.t. the utterance, linear precedence is only physically evaluated, no functional correction is introduced.

3.1 Referential Policies and Algorithms

There are also two general referential policy assumption that we adopt in our approach: The first one is related to pronominal expressions, the second one to referring expressions or entities to be asserted in the History List, and are expressed as follows:

- no more than two pronominal expressions are allowed to refer back in the previous discourse portion;
- at discourse level, referring expressions are stored in a push-down stack according to Persistence principles.

Persistence principles respond to psychological principles and limit the topicality space available to user w.r.t. a given text. It has a bidimensional nature: it is determined both in relation to an overall topicality frequency value and to an utterance number proximity value.

Only "persistent" referring expressions are allowed to build up the History List, where persistence is established on the basis of the frequency of topicality for each referring expression which must be higher than 1. All referring expression asserted as Topic (Secondary, Potential) only once are discarded in case they appeared at a distance measured in 5 previous utterances. Proximate referring expressions are allowed to be asserted in the History List.

In particular, if Mitkov considers the paragraph as the discourse unit most suitable for coreferring and cospecifying operation at discourse level, we prefer to adopt a parameterized procedure which is definable by the user and activated automatically: it can be fired within a number that can vary from every 10 up to 50 sentences. Our procedure has the task to prune the topicality space and reduce the number of perspective topic for Main and Secondary Topic. Thus we garbage-collect all non-relevant entities. This responds to the empirically validated fact that as the distance between first and second mention of the same referring expression increases, people are obliged to repeat the same linguistic description, using a definite expression or a bare NP. Indefinites are unallowed and may only serve as first mention; they can also be used as bridging expression within opaque propositions. The first procedure is organized as follows:

A. For each clause,

1. we collect all referential expressions and weight them (see B below for criteria) – this is followed by an automatic ranking;
2. then we subtract pronominal expressions;
3. at clause level, we try to bind personal and possessive pronouns obeying specific structural properties; we also bind reflexive pronouns and reciprocals if any, which must be bound obligatorily in this domain;
4. when binding a pronoun, we check for disjointness w.r.t. a previously bound pronoun if any;
5. all unbound pronouns and all remaining personal pronouns are asserted as "externals", and are passed up to the higher clause levels;

B. Weighting is carried out by taking into account the following linguistic properties associated to each referring expression:

1. Grammatical Function with usual hierarchy (SUBJ > ARG_MOD > OBJ > OBJ2 > IOBJ > NCMOD);
2. Semantic Roles, as they have been labelled in FrameNet, and in our manually produced frequency lexicon of English;
3. Animacy: we use 75 semantic features derived from WordNet, and reward Human and Institution/Company labelled referring expressions;
4. Functional Clause Type is further used to introduce penalties associated to those referring expressions which don't belong to main clause.

C. Then we turn at the higher level – if any -, and we proceed as in A., in addition

1. we try to bind pronouns passed up by the lower clause levels
   - if successful, this will activate a retract of the "external" label and a label of "antecedenthood" for the current pronoun with a given antecedent;
   - the best antecedent is chosen by recursively trying to match features of the pronoun with the first available antecedent previously ranked by weighting;
   - here again whenever a pronoun is bound we check for disjointness at utterance level.

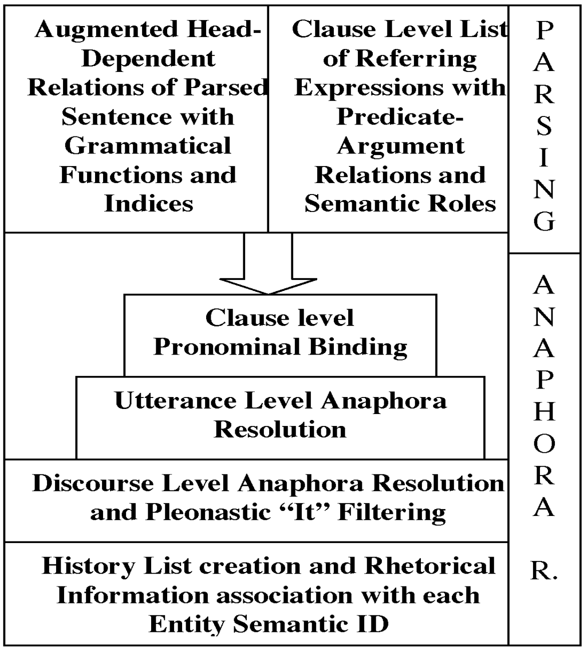
D. This is repeated until all clauses are examined and all pronouns are scrutinised and bound or left free.

E. Pronouns left free – those asserted as externals – will be matched tentatively with the best candidates provided this time by a "centering-like" algorithm.

Step A. is identical and is recursively repeated until all clauses are processed.

Then, we move to step B. which in this case will use all referring expressions present in the utterance, rather than only those available locally.

Fig. 1 GETARUNS AR algorithm


Augmented Head-Dependent Relations of Parsed Sentence with Grammatical Functions and Indices
Clause Level List of Referring Expressions with Predicate-Argument Relations and Semantic Roles
PARSING
Clause level Pronominal Binding
Utterance Level Anaphora Resolution
Discourse Level Anaphora Resolution and Pleonastic "It" Filtering
History List creation and Rhetorical Information association with each Entity Semantic ID
ANAPHORA R.

3.2 Focussing Revisited

Our version of the focussing algorithm follows Sidner's proposal (Sidner C., 1983; Grosz B., Sidner C., 1986), to use a Focus Stack, a certain Focus Algorithm with Focus movements and data structures to allow for processing simple inferential relations between different linguistic descriptions co-specifying or coreferring to a given entity.

Our Focus Algorithm is organized as follows: for each utterance, we assert three "centers" that we call Main, Secondary and the first Potential Topic, which represent the best three referring expressions as they have been weighted in the candidate list used for pronominal binding; then we also keep a list of Potential Topics for the remaining best candidates. These three best candidates repositories are renovated at each new utterance, and are used both to resolve pronominal and nominal cospecification and coreference: this is done both in case of strict identity of linguistic description and of non-identity. The second case may occur either when derivational morphological properties allow the two referring expressions to be matched successfully, or when a simple hyponym/hypernym relation is entertained by two terms, one of which is contained in the list of referring expressions collected from the current sentence, and the other is among one of the entities stored in the focus list.

The Main Topic may be regarded the Forward Looking Center in the centering terminology or the Current Focus. All entities are stored in the History List (HL) which is a stack containing their morphological and semantic features: this is not to be confused with a Discourse Model - what we did in the deep complete system anaphora resolution module – which is a highly semantically wrought elaboration of the current text. In the HL every new entity is assigned a semantic index which identifies it uniquely. To allow for Persistence evaluation, we also assert rhetorical properties associated to each entity, i.e. we store the information of topicality (i.e. whether it has been evaluated as Main, Secondary or Potential Topic), together with the semantic ID and the number of the current utterance. This is subsequently used to measure the degree of Persistence in the overall text of a given entity, as explained below.

In order to decide which entity has to become Main, Secondary or Potential Topic we proceed as follows:

- we collect all entities present in the History List with their semantic identifier and feature list and proceed to an additional weighting procedure;
- nominal expressions, they are divided up into four semantic types: definite, indefinite, bare NPs, quantified NPs. Both definite and indefinite NP may be computed as new or old entity according to contextual conditions as will be discussed below and are given a rewarding score;
- we enumerate for each entity its persistence in the previous text, and keep entities which have frequency higher than 1, we discard the others;
- we recover entities which have been asserted in the HL in proximity to the current utterance, up to four utterances back;
- we use this list to "resolve" referring expressions contained in the current utterance;
- if this succeeds, we use the "resolved" entities as new Main, Secondary, and Potential Topics and assert the rest in the Potential Topics stack;
- if this fails – also partially – we use the best candidates in the weighted list of referring expressions to assert the new Topics. It may be the case that both resolved and current best candidates are used, and this is by far the most common case.

4. Evaluation and General Discussion

Evaluating anaphora resolution systems calls for a reformulation of the usual parameters of Precision and Recall as introduced in IR/IE field: in that case, there are two levels that are used as valuable results; a first stage where systems are measured for their capacity to retrieve/extract relevant items from the corpus/web (coverage-recall). Then a second stage follows in which systems are evaluated for their capacity to match the content of the query (accuracy-precision). In the field of IR/IE items to be matched are usually constituted by words/phrases and pattern-matching procedures are the norm. However, for AR systems this is not sufficient and NLP heavy techniques are used to get valuable results. As Mitkov also notes, this phase jeopardizes the capacity of AR systems to reach satisfactory accuracy scores simply because of its intrinsic weakness: none of the off-the-shelf parsers currently available overcomes 90% accuracy.

To clarify these issues, we present here below two Tables: in the first one we report data related to the vexed question of whether pleonastic "it" should be regarded as part of the task of anaphora resolution or rather part of a separate classification task – as suggested in a number of papers by Mitkov. In the former case, they should contribute to the overall anaphora resolution evaluation metrics; in the latter case they should be compute separately as a case of classification over all occurrences of "it" in the current dataset and discarded from the overall count. Even though we don't agree fully with Mitkov's position, we find it useful to deal with "it" separate, due to its high inherent ambiguity. Besides, it is true that the AR task is not like any Information Retrieval task.

In Table 1 below we reported figures for "it" in order to evaluate the three algorithms in relation to the classification task. Then in Table 2. we report general data where we computed the two types of accuracy reported in the literature. In Table 1 we split results for "it" into Wrong Reference vs Wrong Classification: following Mitkov, in case we only computed anaphora related cases and disregarded those cases of "it" which were wrongly classified as expletives. Expletive "it" present in the text are 189: so at first we computed coverage and accuracy with the usual formula that we report below. Then we subtracted wrongly classified cases from the number of total "it" found in one case (following Mitkov who claims that wrongly classified "it" found by the system should not count; in another case, this number is subtracted from the total number of "it" to be found in the text. Only for MARS we then computed different measures of Coverage and Accuracy. If we regard this approach worth pursuing, we come up with two Adjusted Accuracy measures which are related to the revised total numbers of anaphors by the two subtractions indicated above.

Table 1. Expletive "it" compared results

	MARS	JavaRAP	GuiTAR	GETARUNS
Coverage	163 (86.2%)	188 (99.5%)	188 (99.5%)	171 (91%)
Accuracy 1	63 (33.3%)	73 (38.6%)	75 (39.7%)	87 (46 %)
Wrong Classification	44 163-44=119 189-44=145	49 189-49=140	64 189-64=125	53 189-53=136
Wrong Reference	56	66	49	32
Accuracy 2	63 (38.6%)			
Adjusted Accuracy 2	63 (52.9%)			
Adjusted Accuracy 3	63 (43.4%)	73 (52.1%)	75 (60%)	87 (64 %)

We computed manually all third person pronominal expressions and came up with a figure 982 which is only confirmed by one of the three systems considered: JavaRAP. Pronouns considered are the following one, lower case and upper case included:

Possessives – his, its, her, hers, their, theirs

Personals – he, she, it, they, him, her, it, them (where "it" and "her" have to be disambiguated)

Reflexives – himself, itself, herself, themselves

There are 16 different wordforms. As can be seen from the table below, apart from JavaRAP, none of the other systems considered comes close to 100% coverage.

Computing general measures for Precision and Recall we have three quantities (see also Poesio & Kabadjov):

- total number of anaphors present in the text;
- anaphors identified by the system;
- correctly resolved anaphors.

Formulas related to Accuracy/Success Rate or Precision are as follows: Accuracy1 = number of successfully resolved anaphors/number of all anaphors; Accuracy2 = number of successfully resolved anaphors/number of anaphors found (attempted to be resolved). Recall - which should correspond to Coverage - we come up with formula: R= number of anaphors found /number of all anaphors to be resolved (present in the text). Finally the formula for F-measure is as follows: 2*P*R/(P+R) where P is chosen as Accuracy 2.

Table 2. Overall results Coverage/Accuracy

	COVERAGE	ACCURACY 1	ACCURACY 2	F-measure
MARS	936 (95.3%)	403/982 (41.5%)	403/903 (43%)	59.26%
JavaRAP	981 (100%)	490/982 (49.9%)	490/981 (50%)	66.7%
GUITAR	824 (84.8%)	445/982 (45.8%)	445/824 (54%)	65.98%
GETARUNS	885 (90.1%)	555/982 (56.5%)	555/885 (62.7%)	73.94%

In absolute terms best accuracy figures have been obtained by GETARUNS, followed by JavaRAP. So it is still thanks to the classic Recall formula that this result stands out clearly. We also produced another table which can however only be worked out for our system, which uses a distributed approach. We managed to separate pronominal expressions in relation to their contribution at the different levels of anaphora resolution considered: clause level, utterance level, discourse level. At clause level, only those pronouns which must be bound locally are checked, as is the case with reflexive pronouns, possessives, some cases of expletive 'it': both arguments and adjuncts may contribute the appropriate antecedent. At utterance level, in case the sentence is complex or there is more than one clause, also personal subject/object pronouns may be bound (if only preferentially so). Eventually, those pronouns which do not find an antecedent are regarded discourse level pronouns.

We collapsed under CLAUSE all pronouns bound at clause and utterance level; DISCOURSE contains only sentence external pronouns. Expletives have been computed in a separate column.

Table 3. GETARUNS pronouns collapsed at structural level

	CLAUSE	DISCOURSE	EXPLETIVES	TOTALS
Pronouns found	410	366	109	885
Correct	266	222	67	555
Errors made	144	144	42	330

As can be noticed easily, the highest percentage of pronouns found is at Clause level: this is not however the best performance of the system, which on the contrary performs better at discourse level. Expletives contribute by far the highest correct result. We also found correctly 47 'there' expletives and 6 correctly classified pronominal 'there' which however have been left unbound. The system also found 48 occurrences of deictic discourse bound "this" and "that", which corresponds to the full coverage.

Finally, nominal expressions: the History List (HL) has been incremented up to 2243 new entities. The system identified 2773 entities from the HL by matching their linguistic description. The overall number of resolution actions taken by the Discourse Level algorithm is 1861: this includes both cases of nominal and pronominal expressions. However, since only 366 can be pronouns, the remaining 1500 resolution actions have been carried out on nominal expressions present in the HL. If we compare these results to the ones computed by GuiTAR, which assign semantic indices to NamedEntities disregarding their status of anaphora, we can see that the whole text is made up of 12731 NEs. GuiTAR finds 1585 cases of identity relations between a NE and an antecedent. However, GuiTAR introduces always new indices and creates local antecedent-referring expression chains rather than repeating the same index of the chain head. In this way, it is difficult if not impossible to compute how many times the text corefers/cospecifies to the same referring expressions. On the contrary, in our case, this can be easily computed by counting how many times the same semantic index is being repeated in a "resolution" or "identity" action of the anaphora resolution algorithm. For instance, the Jury is coreferred/cospecified 12 times; Price Daniel also 12 times and so on.

5. Conclusions

The error rate of both Charniak's and Connexor's as reported in the literature, is approximately the same, 20%; this notwithstanding, MARS has a slightly reduced coverage when compared with JavaRAP, 96%. GuiTAR has the worst coverage, 85%. As to accuracy, none of the three algorithms overruns 50%: JavaRAP has the best score 49.9%. However GETARUNS has 63% correct score, with 90% coverage.

There are at least three reasons why our system has a better performance: one is the presence of a richer functional and semantic information as explained above, which comes with augmented head-dependent structures. Second reason is the decision to split the referential process into two and treat utterance level pronominal expressions separately from discourse level ones. Third reason is the way in which discourse level anaphora resolution is organized: our version of the Centering algorithm hinges on a record of a list of best antecedents weighted on the basis of their behaviour in History List and on their intrinsic semantic properties. These three properties of our AR algorithm can be dubbed the Knowledge Rich approach.

F-measures approximates very closely what we obtained in a previous experiment: however, as a whole it is an insufficient score to insure adequate confidence in semantic substitution of anaphoric items by the head of the antecedent. Improvements need to come from parsing and the lexical component.

Acknowledgements

Thanks to three anonymous reviewers who helped us improve the overall layout of the paper.

References

Delmonte R. 1990. Semantic Parsing with an LFG-based Lexicon and Conceptual Representations, *Computers & the Humanities,* 5-6, pp.461-488.

Delmonte R. and D.Bianchi 1991. Binding Pronominals with an LFG Parser, *Proceeding of the Second International Workshop on Parsing Technologies*, Cancun(Messico), ACL 1991, pp.59-72.

Delmonte R., D.Bianchi 1992. Quantifiers in Discourse, in *Proc. ALLC/ACH'92*, Oxford(UK), OUP, pp. 107-114.

Delmonte R. 1992. *Linguistic and Inferential Processing in Text Analysis by Computer*, UP, Padova.

Delmonte R. and D.Bianchi 1994. Computing Discourse Anaphora from Grammatical Representation, in D.Ross & D.Brink(eds.), *Research in Humanities Computing* 3, Clarendon Press, Oxford, 179-199.

Delmonte R. and D.Bianchi 1999. Determining Essential Properties of Linguistic Objects for Unrestricted Text Anaphora Resolution, *Proc. Workshop on Procedures in Discours*e, Pisa, pp.10-24.

Delmonte R., L.Chiran, and C.Bacalu, (2000). Towards An Annotated Database For Anaphora Resolution, LREC, Atene, pp.63-67.

Delmonte R. 2002a. From Deep to Shallow Anaphora Resolution: What Do We Lose, What Do We Gain, in Proc. International Symposium RRNLP, Alicante, pp.25-34.

Delmonte R. 2002b. From Deep to Shallow Anaphora Resolution:, in *Proc. DAARC2002 , 4th Discourse Anaphora and Anaphora Resolution Colloquium*, Lisbona, pp.57-62.

Delmonte, R. 2003. Getaruns: a Hybrid System for Summarization and Question Answering. In Proc. Natural Language Processing (NLP) for Question-Answering, EACL, Budapest, pp. 21-28.

Delmonte R. 2004. Evaluating GETARUNS Parser with GREVAL Test Suite, In Proc. ROMAND - 20th COLING, University of Geneva, pp. 32-41.

Di Eugenio B. 1990. Centering Theory and the Italian pronominal system, COLING, Helsinki.

Grosz B. and C. Sidner 1986. Attention, Intentions, and the Structure of Discourse, *Computational Linguistics* 12 (3), 175-204.

Kennedy, C. and B. Boguraev, 1996. Anaphora for everyone: Pronominal anaphora resolution without a parser. In *Proc. of the 16th COLING*, Budapest.

Long Qiu, Min-Yen Kan, and Tat-Seng Chua, 2004. A Public Reference Implementation of the RAP Anaphora Resolution Algorithm, In *Proceedings of the Language Resources and Evaluation Conference 2004 (LREC 04),* Lisbon, Portugal, pp.1-4.

Mitkov R. 1995. Two Engines are better than one: Generating more power and confidence in the search for the antecedent, *Proceedings of Recent Advances in Natural Language Processing*, Tzigov Chark, 87-94.

Mitkov, R. 1998. Robust Pronoun Resolution with limited knowledge. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING'98)/ACL'98 Conference*, pp. 869-875, Montreal, Canada.

Mitkov, R., R. Evans, and C. Orasan. 2002. A New, Fully Automatic Version of Mitkov's Knowledge-Poor Pronoun Resolution Method, *Proceedings of CICLing*-2002, pp.1-19.

Poesio, M. and R. Vieira, 1998. A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216.

Poesio, M. and Mijail A. Kabadjov 2004. A General-Purpose, off-the-shelf Anaphora Resolution Module: Implementation and Preliminary Evaluation *Proceedings of the Language Resources and Evaluation Conference 2004 (LREC 04),* Lisbon, Portugal, pp.1-4.

Sidner C. 1983. Focusing in the Comprehension of Definite Anaphora, in Brady M., Berwick R.(eds.), *Computational Models of Discourse*, MIT Press, Cambridge, MA, 267-330.

Webber B. 1983. So can we Talk about Now?, in Brady M., Berwick R.(eds.), *Computational Models of Discourse*, MIT Press, Cambridge, MA, 331-371.

Webber B. L. 1991. Structure and Ostension in the Interpretation of Discourse Deixis, in *Language and Cognitive Processes* 6 (2):107-135.

Zribi-Hertz A. 1989. Anaphor Binding and Narrative Point of View: English reflexive pronouns in sentence and discourse, *Language*, 65(4):695-727.