AR-Engine – a framework for unrestricted co-reference resolution

Dan Cristea*, Oana-Diana Postolache*, Gabriela-Eugenia Dima■, Cătălina Barbu□

*University "Al.I.Cuza" of Iaşi, Faculty of Computer Science, 16 Berthelot St., 6600 Iaşi, Romania
{dcristea, oanap}@infoiasi.ro
■University "Al.I.Cuza" of Iaşi, Faculty of Letters, 11, Carol I Bvd., 6600 Iaşi, Romania
g.dima@mail.uaic.ro
□University of Wolverhampton, School of Humanities, Languages and Social Sciences,
Stafford St., WV1 1SB, Wolverhampton, United Kingdom
C.Barbu@wlv.ac.uk

Abstract

The paper presents a framework that allows the design, realisation and validation of different anaphora resolution models on real texts. The type of processing implemented by the engine is an incremental one, simulating the reading of texts by humans. Advanced behaviour like postponed resolution and accumulation of values for features of the discourse entities during reading is implemented. Four models are defined, plugged in the framework and tested on a small corpus. The approach is open to any type of anaphora resolution. However, the models reported deal only with co-reference anaphora, independent of the type of the anaphor. It is shown that the setting on of more and more features, generally results in an improvement of the analysis.

1. Introduction

It is well known that an algorithm of anaphora resolution (AR) with a very high degree of success has not been found yet. In (Cristea and Dima, 2000) a framework able to easily accommodate different approaches in anaphora resolution was proposed. The central notion in the framework is that of **anaphora resolution model**. In this paper we describe a group of experiments with the framework, used as a workbench, and show an ascending precision and recall series of results that are obtained by AR models that have more and more features turned-on. The experiments report results on co-reference resolution that involves any type of noun phrases as anaphors. The work described here relies heavily on annotated language resources.

In section 2 we describe a framework that allows easy design, implementation and evaluation of any AR model. The section 3 presents the small corpus we used for our research and the experiments pursued. Section 4 illustrates a series of four models that were used for the experiments and section 5 reports the results.

2. A framework for anaphora resolution

In (Cristea and Dima, 2000) the anaphora resolution process is interpreted as involving three layers: the **text layer** – populated with referential expressions (REs) –, the **intermediate layer** – where feature structures (FSs) are filled-in with information from the text –, and the deep **semantic layer** – where descriptions of discourse entities (DEs) are placed. We say that an FS is **projected** from an RE and that a DE is **evoked** by an FS (Figure 1).

The type of analysis supported by the framework is incremental; therefore the order of processing the anaphors simulates the human reading. Just like in normal reading, anaphors are mostly resolved at the time of reading, but sometimes decisions are postponed until the acquisition of complementary information that helps the disambiguation process. It is like when backwards eye movements reveal indecisions (Vonk, 1985).

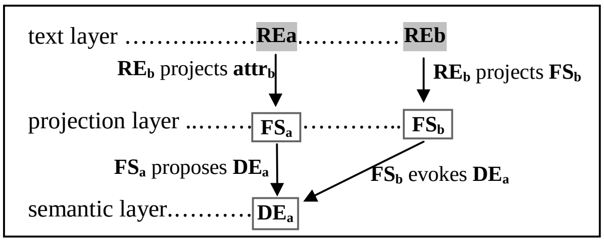

Figure 1. Three-layer representation of co-referring referential expressions

2.1. Definition of an AR model

In the definition of an AR approach the notion of **AR model** is basic. In (Cristea and Dima, 2001) such a model is described as having four components:

- a **set of primary attributes** that fill the projection layer and are then reported to the semantic layer;
- a **set of knowledge sources** or virtual processors fetching values for the attributes during the incremental text processing;
- a **set of heuristics** or rules intended to co-operate in order to decide if an RE introduces a new discourse entity and, if not, what existing DEs it co-refers;
- a set of rules that configure the **domain of referential accessibility**, therefore establishing the order in which DEs have to be checked.

In order for the framework to accommodate an AR approach, a specific model has to be plugged in the framework. The framework, at least in principle, allows for unrestricted AR and only the plugged-in model dictates whether the resolution involves only co-references or any type of anaphora and whether it takes into consideration only pronouns or any type of referential expressions.

An analysis of the existent approaches (Mitkov, 2002) helps classify the attributes into the following categories:

a. morphological:

- lexical number;
- lexical gender;
- person.

All the approaches use morphological criteria to filter out antecedents. However, the elimination of possible referential candidates based on mismatches in morphological features may lead to errors as shown at least by Barlow (1998). Morphology cannot be ignored, but a less categorical approach is preferable.

b. syntactical:

- full syntactic description of REs as constituents of a syntactic tree (Lappin and Leass, 1994);
- marking of the syntactic role for subject position or obliqueness (the subcategorisation function in respect to the verb) of the REs: syntactic domain-based approaches (Chomsky, 1981; Reinhart, 1981; Gordon and Hendricks, 1998; Kennedy and Boguraev, 1996) and all Centering-based approaches (Grosz, Joshi and Weinstein, 1995; Brennan, Friedman and Pollard, 1987);
- quality of being adjunct, embedded or complement of a preposition (Kennedy and Boguraev, 1996);
- inclusion or not in an existential construction (Kennedy and Boguraev, 1996);
- syntactic patterns in which the RE is involved, that can lead to the determination of syntactic parallelism (Kennedy and Boguraev, 1996; Mitkov, 1997).

c. semantic:

- position of the head of the RE in a conceptual hierarchy (hypo/hypernymy): all models using Wordnet (Poesio, Vieira and Teufel, 1997). Features as animacy, sex (or natural gender) and concreteness could be considered simplified semantic tags derived from a conceptual hierarchy;
- inclusion in a synonymy class that is determined by the context;
- semantic roles, out of which selectional restrictions, inferential links, pragmatic limitations, semantic parallelism and object preference can be verified.

d. positional:

- offset of the first token of the RE (a noun phrase – NP) in the text (Kennedy and Boguraev, 1996);
- inclusion in an utterance, sentence or clause, considered as a discourse unit (Azzam, Humphreys and Gaizauskas, 1998). The setting of this feature allows the computation of the proximity between the anaphor and the antecedent. If both the anaphor and the antecedent are placed in the same unit than c-commands criteria can be applied; if they are placed in different units, then the number of units in between could be used as a criterion in limiting the domain of referential accessibility (see Component 4 below).

e. surface realisation (type):

- zero-pronoun (also called zero-anaphora or non-text string), clitic pronoun, full pronoun, reflexive pronoun, possessive pronoun, demonstrative pronoun, reciprocal pronoun, expletive “it”, bare noun (undetermined NP), indefinite NP, definite NP, proper noun (name).

f. others:

- inclusion or not of the RE in a specific lexical field, dominant in the text (called *domain concept* in Mitkov, 1997);
- frequency of the term in the text (Mitkov, 1997);
- occurrence of the term in a heading (Mitkov, 1997).

Component 2 includes **a set of knowledge sources** fetching values for the attributes of the feature structures belonging to the projection layer.

What we understand by a knowledge source is a virtual processor able to fill in values for one single attribute on the projection layer, for instance lexical number, or lexical gender, or part of speech, or syntactic role. Practically, current processors simultaneously fetch values for more than one such attribute. Thus, a morpho-syntactic tagger represents several knowledge sources as it fills in values for more than one attribute of the head word of the RE (Tufis, 2000).

At least two knowledge sources are fundamental to all possible models of anaphora resolution: a part of speech tagger, to associate tags to every token of the text, and an NP extractor, capable of recognising REs by grouping text tokens into NPs. These should be structured as head-modifier compounds. As there are NPs which cannot be REs (i.e. *the bucket* within the verbal phrase *to kick the bucket*), a parser should reject such noun phrases, by running a set of regular expressions to discover phrasal units.

Besides the two processors above, currently existing systems use additional knowledge sources. For instance, Kennedy and Boguraev (1996) introduce a marker of syntactic function and a set of patterns which recognise the expletive “it” (near specific sets of verbs or as subject of adjectives with clausal complements). Azzam, Humphreys and Gaizauskas (1998) use a syntactic analyser, a semantic analyser, and an elementary event finder. Gordon and Hendrick (1998) employ a surface realisation identifier and a syntactic parser, while Hobbs (1978) requires, for his semantic approach, a syntactic analyser, a surface realisation identifier and a set of axioms to determine semantic roles and relations of lexical items.

Component 3 contains **a set of rules and heuristics** intended to answer one or both of the following two questions involving the current FS: (1) Does it introduce a new discourse entity? (2) Which one (ones) of the existing DEs does it refer to and what is the referential relation between this FS and the respective DE (DEs), considered as antecedent(s)? The rules/heuristics perform the evoking phase between an existent FS belonging to the projection layer and DEs belonging to the semantic layer. The mention of more than one DE in the second point above is due to the observation that there are cases when an FS (and, therefore, its corresponding RE) could be in more than just one referential relation to already identified DEs. In the sequence *John … his house*, for instance, there are three REs – *John, his house*, and *his* – and *his* is in a co-referential relation with the entity [**John**] while being in an ownership relation with [**his house**].

In accordance with most authors, we accomplish this process by two types of rules:

- certifying rules (applied first), which, if evaluated to true on a pair (FS, DE), certify without ambiguity the DE as an antecedent of the FS;
- demolishing rules (applied after the certifying rules), which rule out a possible DE as candidate of an FS (and its corresponding RE). These rules lead to a filtering phase that eliminates from among the candidates those discourse entities that cannot possibly be referred to by the RE under investigation;
- promoting/demoting rules (applied after the demolishing rules), which increase/decrease a **resolution score** associated with a pair (FS, DE). These rules allow for a subsequent selection phase, in which either the best candidate of an FS is chosen from the ones remaining after the demolishing rules have been applied, or a new entity is introduced.

To refer again to the known approaches, in order to rule out possible candidates, Kennedy and Boguraev (1996), for instance, implement conditions that prevent a pronoun to co-refer a constituent (NP) which contains it. Thus, in *the child of his brother*, *his* is neither *child*, nor *brother*, but a different entity. For the remaining candidates, they compute the salience by weighing a set of attribute-values pairs. The weights are linguistically and experimentally justified (Keenan and Comrie, 1977; Lappin and Leass, 1994). Gordon and Hendricks (1997) show that the antecedent's syntactic prominence (notion related to the relative distance in a syntactic tree) influences the selection of the co-referential candidate. In (Gordon and Hendricks, 1998) the salience of the relations between names and pronouns is calculated by using a gradation of surface realisation pairs: name-pronoun > name-name > pronoun-name. Therefore, if the surface realisation of the anaphor is a name, then a candidate whose surface realisation is a name will weigh higher than one whose surface realisation is a pronoun.

Apart from these types of rules, the builder of the model can also choose to implement **heuristics** intended to deal with still unresolved (postponed) candidates, in the preceding FSs. The idea is that, at a certain moment during processing, the resolution of new anaphors could bring new proofs for the disambiguation of old, unresolved ones. The processing moments for reconsidering postponed FS resolution, once a new link has been added, varies from the processing moment of the immediately following RE in the text, up to the processing moment of the last RE belonging to the following discourse unit. We haven't found cases when this reconsideration should be done later than this moment. At the same time, it is not clear yet whether an untying mechanism should be incorporated into the framework in order to reconsider the anaphoric "garden paths".

Component 4 contains **a procedure that configures the domain of referential accessibility**. This component has two tasks: to establish the semantic DEs that are open to be referred to by a certain RE (forming the domain of referential accessibility – DRA) and to establish their order according to some prominence criteria.

To exemplify the first task, both the attentional state theory (Grosz and Sidner, 1986) and the veins theory (Cristea, Ide and Romary, 1998) discuss accessibility constraints that rely on the discourse structure. By applying these constrains to the position of a certain RE, some DEs from among those already found as resolution candidates could be hidden. In the attentional-based approaches, the accessibility in the current discourse unit is given by the top-down order of states in a focus stack. In the veins theory, the DRA of a unit is an initial part of a sequence of discourse units, called vein, computed from the discourse structure, considered a nucleus/satellite tagged binary tree.

For the second task, different strategies could be imagined. The recency order is extensively used, placing all the discourse entities referred from the REs in the DRA in the most-recent-first order and eliminating duplicates (if different REs refer the same DE, only the most recent reference is kept, in the corresponding place of the list). Following this order, among the set of discourse entities, with respect to which the current anaphor displays the same resolution score, the one closest to it is selected.

For instance, a combination of a "no restriction" DRA policy with the recency order gives the well-known linear order. Alternatively, focus-based approaches (Sidner, 1981; Azzam, Humphreys and Gaizauskas, 1998) use registers for current focus, and alternate foci list, which are updated after each sentence, and, based on them, define an order in which to look for the antecedent. In their discourse prominence representation approach, Gordon and Hendrick (1998) speak about "ordering of entities in the discourse order that determines the accessibility of those entities as referents for subsequent expressions". In VT the E-DRA of a discourse unit is the hierarchical DRA taken from the unit's vein plus the extra units, taken as the least prominent. The order of DEs as possible antecedents of a given RE used in VT-like approaches is the recency order of DEs corresponding to REs belonging to this list (therefore a combination of hierarchical E-DRA with the recency ordering).

2.2. Processing in the framework

The framework processes the text sequentially. When processing begins, the text layer contains the original text, with the sequence of REs, there is no FS built on the projection layer and the semantic layer is empty of DEs. As the projection layer is only meant to support REs under resolution, no structure is maintained on this layer for the resolved REs. On the contrary, the semantic layer maintains structures corresponding to all already found DEs and is continuously updated with the information contributed by the processed co-referential FSs. At the same time, these DEs define equivalence classes among REs of the text layer, as all co-referential expressions refer to the same DE. There is one link that is established between each RE and a specific DE, as in (Kennedy and Boguraev, 1996), and one DE points back to all REs in its class. As already mentioned, one DE accumulates all the relevant properties of its REs, for instance: natural gender and number, all lemmas and proper names that were used to refer that entity. This is why the evoking rules usually match gender and number by verifying inclusion of the FSs values in the corresponding attribute values of the DE, instead of performing an equality check.

The moment an RE_x is read, or shortly afterwards (to enable the existent knowledge sources to acquire enough information), a representation on the projection layer is built – the **FS$_x$**. The processing moment of an RE may be delayed in case some knowledge sources require more text than the span of tokens included in the RE itself or in the immediately preceding context. For instance, some semantic features of REs are identifiable only after the processing of the verb they relate to, while zero pronouns can be identified only after the identification of the verb.

When all the declared knowledge sources in the second component of the model contributed with their information, the recently build FS_x on the projection layer has a filled-in list of attribute-value pairs: $a_i = v_i$, for all attributes listed by the first component of the model. Suppose DRA=(DE_1, … DE_m) is the ordered list of accessible discourse entities at the current moment, as given by the fourth component of the model, with DE_1 – the most prominent. Let the three set of rules be denoted as follows: certifying rules – *CR*, demolishing rules – *DR* and scored rules – *SR* and the heuristics be denote by *H*. Among the attributes of the FS_x, a *candidates* attribute is a vector of pairs, where in each pair the first field points to a corresponding DE, and the second field is the resolution score of the current anaphor RE_x with regard to the corresponding DE as a possible antecedent. Let's denote the two slots of the elements in this list as *idx* and *score*, respectively. Then, the evoking phase of a co-reference type resolution corresponding to FS_x and a given *DRA* list runs as follows:

```
procedure evoke($FS_x$, DRA)
1.  initialise the candidates list of $FS_x$
      with one element for each discourse
      entity $DE_j$ in DRA as follows:
        candidates.idx := a pointer to a $DE_j$,
          in the order given in DRA;
        candidates.score := 0;
2.  for each rule $r \in CR$ and each $DE_j \in DRA$ do{
      if $r(FS_x,DE_j)$ = true then {
        assign $DE_j$ as the antecedent of $RE_x$;
        go to 9;}
3.  for each rule $r \in DR$ and each $DE_j$ do{
      if $r(FS_x,DE_j)$ = true then
        eliminate $DE_j$ from the candidates
          list of $FS_x$;}
4.  for each rule $r \in SR$ and each remaining
      $DE_j$ in the candidates list of $FS_x$ do{
        candidates.score := $\frac{\sum_{r \in SR} w_r \times s_r}{\sum_{r \in SR} w_r}$
        where $w_r$ is the weight of rule $r$ and
          $s_r$ is the matching score of $r$
          applied to the pair ($FS_x$,$DE_j$);}
5.  sort the candidates list in the
      descending order of the score values
      and then of the idx values;
6.  if candidates.score(0) < $threshold_{min}$
      then {
        copy $FS_x$ as $DE_{m+1}$ and connect the
          current anaphor ($RE_x$) with it;
        go to 10;}
7.  else if within the $threshold_{diff}$ range
      of values there is more than one
      candidates.score then {
        keep in the candidates list of $FS_x$
          only the entries corresponding to
          this range of scores and delete the
          rest;
        return false;}
8.  else choose as antecedent of $RE_x$ the $DE_j$
      given by candidates.idx(0), i.e. the
      first ranked candidate after sorting;
9.  replace $DE_j$ with the merge between $FS_x$
      and the previous content of $DE_j$;
10. delete $FS_x$ from the projection layer;
11. for each $FS_y$ remained on the projection
      layer do{
        if evoke($FS_y$, DRA)=false, where now
          DRA equals the corresponding
          candidates.idx list, then apply
          $h(FS_y, DRA)$, where $h \in H$;}
12. return true;
```

Steps 2, 3 and 4 describe actions of certifying rules, demolishing rules and scored rules, respectively. If a certifying rule fired then the antecedent is found and the procedure continues with step 9. If a demolishing rule fires, the targeted DE is eliminated from the candidates. The remaining candidates are then sorted at step 5 in the descending order of the resolution scores computed by the scored rules. Step 6 describes the actions to be taken when a new discourse entity is proposed, due to poor matching of the projected features with any of the already existent DEs. Sometimes, two or even more structures could be maintained on the projection layer, as revealed at step 7. This happens in the case of postponed resolution, which is triggered by a too small ranking difference between the best-ranked candidates. The threshold values used in steps 6 and 7 ($threshold_{min}$ and $threshold_{diff}$) are included in the model among other fine tuning parameters. If just one DE neatly differentiates from the others of the DRA, then at step 8 it is taken as the antecedent. In all cases when an antecedent is found (at step 2 – as a result of a certifying rule or at step 8 – by finding a well individualised candidate among others), the current FS injects information into the existent DE structure – at step 9. This is the way in which we allow the DEs to evolve along with the unfolding of the discourse. Then the current FS, successfully resolved, may disappear from the projection layer – at step 10. Finally, at step 11, which is reached just in case the current FS is resolved, the remaining unsolved FSs on the projection layer are again considered. There are two ways in which unsolved FSs can benefit from FSs resolved at a later point in the discourse. One is by running again the set of rules for just those candidates kept in its list. Since resolved FSs merge their list of attribute-values with the one of the assigned DE, it is possible that these new pieces of information injected into DEs give the illuminating clue to help the disambiguation in a case of indecision. For instance, in the sequence:

When Roberta came home Emily was studying.
She really wanted to improve her grades.

the moment the pronoun *she* can be interpreted is not sooner than the end of the whole sentence. Isolated from the following context, *she* is ambiguous between [**Roberta**] and [**Emily**]. A semantic restriction prefers the same person that is studying to improve her grades, which results in linking *her* to [**Emily**]. Then, it is more likely that, if someone tries to improve someone's grades, than the two persons be the same, resulting in *she* being linked to the same [**Emily**]. If the second sentence would have been:

She really wanted to convince her to go out.

then a similar inference scheme would first have to find *her* as being [**Emily**], because the person that needs to be convinced to go out is rather someone who is busy studying than someone who just comes home, and then a syntactic restriction would prefer *she* to be [**Roberta**] on the ground that if they were the same person, the pronoun *herself* should have been used.

The framework is, in itself, language independent. The adjustment to one language or another is done by setting the list of attributes to the desired language, defining the knowledge sources capable to fill them and applying evoking rules/heuristics specific to each language. There is no reason to believe that the criteria which define the domain of accessibility would be language specific.

3. The method

For our experiments we have used a small corpus from Orwell's novel "Nineteen eighty four"[1]. The golden standard included the FDG automatic parsing, manually corrected, and a manual annotation to co-references. The annotation format was XML (a Perl script translated the FDG output to XML). For co-reference annotation, right-to-left linear chains were preferred to references targeting the first appearance of a discourse entity in the text. All referential expressions that are textually realised by noun phrases resulted as a by-product of the FDG annotation, therefore we have not restricted the research only to pronouns.

To evaluate the approach, we've tracked five characters from this short fragment (Winston Smith, coded as [**Winston**], the girl with black hair – [**the girl**], the sand-like haired woman – [**the woman**], O'Brien – [**O'Brien**] and Goldstein – [**Goldstein**]). We have experimented four models, one of them with two variations, that we believed should have given an increasing degree of resolution, as they used more and more features turned-on, therefore involving more and more computational power. For each co-referential anaphoric chain and for each model, we've computed the precision and recall.

4. The models used in the experiment

The first model (**Base-model**) is described by the following quadruple:

- set of attributes: POS (part of speech), LNUM (lexical number), LGEN (lexical gender), LEM (lexical lemma), NAME (proper name – if it is the case);
- set of knowledge sources: a POS-tagger (contributing values to POS, LNUM, LGEN) and a lemmatiser (contributing values to LEM and NAME);
- set of matching rules:
  - o certifying rules (in what follows *fs* and *de* are an FS and a DE structure that are matched by the rules):
    - ▪ *match-NAME*(*fs*, *de*): if *fs*.POS=proper-noun and *de*.NAME ≠ null and *fs*.NAME ∈ *de*.NAME then true else false;
  - o scored rules:
    - ▪ *match-LNUM*(*fs*, *de*): if *fs*.LNUM ≠ null and *de*.LNUM ≠ null and *fs*.LNUM ∈ *de*.LNUM then 1 else 0;
    - ▪ *match-LGEN*(*fs*, *de*): if *fs*.LGEN ≠ null and *de*.LGEN ≠ null and *fs*.LGEN ∈ *de*.LGEN then 1 else 0;
    - ▪ *match-LEM*(*fs*, *de*): if *fs*.POS=common-noun and *de*.LEM ≠ null and *fs*.LEM ∈ *de*.LEM then 1 else 0;
- domain of referential accessibility: linear recency (look linearly back for matching antecedents within a parametrized range of DEs).

Model 2 (**WN-model**):

- set of attributes: same as in model 1, plus natural gender (shared among three attributes SHE, HE IT, with a score attributed to each of them summing up to 100%), and information related to synonymy and lexical ontology (hypernymy) associated with a noun: the attributes HYPER and SYNO;
- set of knowledge sources: same as in the base-model plus a Wordnet processor able to fetch the synsets of a noun and the respective hypernym ontology as well as the values feminine/masculine/neuter for the natural gender. This source looks for hypernymic concepts of female, male and person in order to distribute scores to the attributes SHE, HE, IT. These scores are computed as follows: any occurrence of the Wordnet concept `<female, female person -- (a person who belongs to the sex that can have babies)>` in a hypernymic chain of a synset of the target word contributes with a SHE vote. Any occurrence of the Wordnet concept `<male, male person -- (a person who belongs to the sex that cannot have babies)>` in a hypernymic chain of a synset of the target word contributes with a HE vote. All the hypernymic chains of the target word that did not match neither the first criterion, nor the second, but meets an occurrence of the Wordnet concept `<person, individual, someone, somebody, mortal, human, soul -- (a human being)>` contributes with a SHE half-a-vote and a HE half-a-vote, and finally all hypernymic chains of the target word that did not match any of the above criteria, contribute with an IT vote. Then all scores are normalized in the range 0-1 by dividing them with the total number of votes. A knowledge source fetching values for natural number was not yet implemented in this model;
- set of matching rules:
  - o certifying rules: *match-NAME* – same as in the base model;
  - o scored rules: *match-LNUM*, *match-LGEN*, *match-LEM* – same as in the base model, then:
    - ▪ *match-NGEN*(*fs*, *de*): true if that attribute in the set SHE, HE, IT in *fs* that has the maximum value corresponds to that with the maximum value in DE, otherwise false;

[1] Four paragraphs from part I, chapter I (924 words).

- *match-SYNO*(*fs*, *de*): true if among the synsets accumulated in the SYNO attribute of the *fs* there is one that belongs also to the set of synsets corresponding to the DE attribute SYNO;
- *match-HYPER*(*fs*, *de*): true if a synset from SYNO belongs to the SYNO attribute of DE or vice-versa;

- domain of referential accessibility: linear, as in the base model.

Model 3 (**Centering-model**): supplementary to the preceding model, this model applies Centering Theory (Grosz, Joshi, Weinstein, 1995; Brennan, Friedman, Pollard, 1987) in order to disambiguate the postponed references, i.e. there are chosen those particular pairs anaphor-antecedent that result in the smoothest transition between adjacent utterances. The Centering-model is composed of:

- sets of attributes: same as in the WN-model plus an attribute SEG that gives the ID of the segment to which the current RE belongs and SYN – keeping its syntactic role as a constituent near the main verb;
- knowledge sources: same as in the WN-model plus FDG (a functional dependency grammar parser capable to fetch syntactic roles);
- set of matching rules and heuristics: same as in the WN-model plus two heuristics:
  - *CT-preference*(`fs,DRA`): among the DEs in *DRA* (postponed *candidates* list of *fs*) prefers that one that is syntactically best placed (subject>direct-object>indirect-object>other) in the preceding unit. Only when two or more members of the *candidate* list of the current FS are referred by REs belonging to the preceding discourse unit this heuristic is activated;
  - *C-command*(*fs*, *DRA*): this heuristic prevents two REs belonging to the same discourse unit and having the subject and direct object positions to share the same referent;
- domain of referential accessibility: linear recency.

Placing Centering preference as a heuristic implements a “default” vision on its applicability: it works only if any other constraints do not result in determining a certain candidate.

Model 4 (**VT-model**): supplementary to the preceding model, this model applies Veins Theory (Cristea, Ide and Romary, 1998) in order to determine a better domain of referential accessibility.

- sets of attributes, knowledge sources and set of matching rules and heuristics – same as in the Centering-model;
- domain of referential accessibility: the combination of hierarchical E-DRA of VT with the recency ordering.

5. Results and conclusions

The easiness to configure AR models within the presented framework allowed experiments intended to verify our initial hypothesis: models 1 to 4, adding more and more features and knowledge sources and being richer and richer in matching rules, as well as in the criteria for defining the domains where to look for antecedents, do they lead to increasing factors of precision and recall? When this prediction does not match the reality, what are the reasons that led to this discrepancy? What conclusions can be drawn? Could the models be fine-tuned to repair the failures?

The most important conclusion seems to be the possibility to combine different knowledge sources, from different approaches, for which to design adequate matching rules and heuristics, therefore to take what is best of all studied models to the benefit of a better AR behaviour. Except for VT related sources of model 4, that need discourse structure annotation and which couldn’t be provided automatically yet, all sources involved in the experiments were automatic, although manually revised.

To a great extend, the results proved the initial hypothesis, namely models behaved better and better as more features were fired. The best results proved 100% precision and values of recall in the range 70% to 100%. Nevertheless, these figures should be taken with care, because of the small dimension of the corpus we could use. So, the lengths of the reference tracks for the five characters in the golden corpus were as follows: [**Winston**] – 23, [**the girl**] – 14, [**the woman**] – 3, [**O’Brien**] – 25 and [**Goldstein**] – 16.

	Model 1		Model 2		Model 3	
	P	R	P	R	P	R
[**Winston**]	100	92	100	92	100	100
[**the girl**]	100	70	100	90	100	90
[**the woman**]	100	33	100	66	100	66
[**O’Brien**]	81	72	81	72	90	72
[**Goldstein**]	100	50	100	50	100	50

Table 1: Precision and recall of models 1 to 3 applied to the five characters tracked

Table 1 displays only the results of the first three models because model 4 did not show any difference from model 3. These results are commented below.

Model 1 resolves all pronouns-against-pronouns references and pronouns-against-proper nouns references (*he* against a DE filled-in with only the information contributed by another *he* pronoun, or *he* against a *Winston* contributed DE). The smaller recall figures for [**the girl**] are due to failure to solve *she* against a DE contributed by the common noun *the girl*, as the Base-model does not have the knowledge that [**the girl**] has a feminine natural gender. As seen by the precision figure less then 100%, some of the references the model predicts for [**O’Brien**] are erroneous. They will be corrected in model 3 only when Centering will be applied. Of the three references for [**the woman**] two fail. One is due to a *who* pronoun, as the model is unprepared yet to treat *who* and *whom*. More interesting is the reference *woman* to a DE contributed by the noun phase *the little sandy-haired woman*. This fails too although a lemma rule is implemented that checked the lemma equality of the head words but its weight is let small to prevent, for instance, all rooms that are referred to in a story to be merged into the same equivalence class. The small recall for [**Goldstein**] in all models is explained because he is referred to as *the Enemy*, *the renegade*, *backslider*, *one of the leading figures of the Party*, *the primal traitor*, *the earliest defiler of the Party’s purity*, etc.

In Model 2 the recall figure for [**the girl**] is improved because now Wordnet knowledge allows linking *she* to a DE contributed by *the girl* as well as *a girl* against *she* as in *She was a bold-looking girl*. This is possible because in 5 cases out of 6 the word *girl* is part of a synset which has as a hypernym the concept female and in the remaining case the concept person is present. This makes the SHE score of the DE [**the girl**] be very high and therefore a match between the anaphor *she* (which sets SHE to 1 and HE and IT to 0) and this DE be high also. Note that this matching rule uses Wordnet to compute a kind of average natural gender out of all senses of the target word, without trying to disambiguate its senses, in the context. In the case of natural gender, an integration of all senses of a word in the attempt to compute an average behavior seems to work pretty well, but in other cases, it proved to yield unacceptable matches. *Woman* against *woman* now also succeeds on the combined ground of equal lemmas and same natural gender.

Initially, Model 2 had another variant in which the synonymy and hypernymy relations were used in order to identify co-reference based on synonymy and on the ontology given by Wordnet concepts. The experiments proved however that, without a sense disambiguation knowledge source, there is little chance that coherent resolutions of the kind *hope – belief* be found without adding many other accidental equivalences as well, as here:

DE59----->RE965: *the impression of being more dangerous than most* | RE977: *The idea* | RE1208: *his mind* | RE1002: *only a dim idea of its nature* | RE1021: *manner* | RE1219: *O'Brien's urbane manner* | RE1053: *a secretly held belief* | RE1220: *a belief* | RE1055: *a hope* | RE1223: *intelligence* | RE1079: *this guess* | RE1146: *death* |

Model 3: In our small corpus there have been found only two cases that took advantage of the Centering heuristic:

> *Winston had seen O'Brien perhaps a dozen times in almost as many years. He felt deeply drawn to him, …*

In this excerpt there is no other clue that *he* refers [**Winston**] and not [**O'Brien**] but the one given by Centering. Indeed if *he* would be [**Winston**] than the transition between the two adjacent units would be CONTINUE, while if *he* would be [**O'Brien**] the transition would be RETAINING and the theory claims that CONTINUE is smoother than RETAINING and, therefore, preferable to it. The *CT-preference* heuristic successfully links *he* to [**Winston**]. Then the *C-command* heuristic hinders *him* to refer to the same entity as *he* does (a direct object is not allowed to refer a subject) and this restricts him to refer [**O'Brien**].

Still, in the example:

> *It was a gesture which, if anyone had still thought in such terms, might have recalled <an eighteenth-century nobleman> offering <<his> snuffbox>*

his correctly refers [**an eighteenth-century nobleman**] because the *C-command* heuristic does not apply. Two new cases were resolved by model 3 with respect to model 2.

Model 4: no long distance references were contained in our small corpus and therefore the application of the VT didn't bring any difference between models 3 and 4.

The following cases are still unresolved: *O'Brien* against *a man called O'Brien* (lack of insight into the constituents of a noun phrase), *one of them* against *a girl* as in *one of them was a girl* (no rule to treat subject – predicative noun relations in nominal predicate constructions and no implemented rule to treat element-to-set relationships), *O'Brien* against *a large, burly man* as in *O'Brien was a large, burly man* (same reason).

Difficulties in obtaining a corpus tagged for co-references prevented us from testing the models on a large corpus. We shall pursue this line in further research, trying also to build and test models for different languages. The AR-Engine that the framework incorporates allows for an easy integration of features that are reported in other approaches. Supplementary, although faithful to an incremental type of processing, the engine allows postponement of resolution until relevant information is acquired. It is also able to accumulate values for features of the discourse entities as the text unfolds.

Acknowledgments. The authors would like to thank Constantin Orasan for providing us with the FDG analysis and NP extraction and Vlad Ciubotariu for the program that helped to perform the manual co-reference annotation.

6. References

Azzam, S., Humphreys, K., and Gaizauskas, R. 1998. Evaluating a Focus-Based Approach to Anaphora Resolution. *Proceedings of the 17th Coling and the 36th Annual Meeting of the ACL (COLING-ACL'98)*. Montreal, Canada.

Barlow, M. 1998. Features Mismatches and Anaphora Resolution. In Botley and T. McEnery (eds). *New Approaches to Discourse Anaphora*. S. Technical Papers Vol 11.

Brennan, S.E., Friedman, M.E., and Pollard, C.J. 1987. A Centering Approach to Pronouns. *Proceedings of the 25th Annual Meeting of the ACL*, Stanford

Chomsky, N. 1981. Lectures on Governement and Binding. Dordrecht, the Netherlands Foris Publishers.

Cristea, D., Ide, N., and Romary, L. 1998. Veins Theory: A Model of Global Discourse Cohesion and Coherence, *Proceedings of the 17th Coling and the 36th Annual Meeting of the ACL (COLING-ACL'98)*. Montreal, Canada.

Cristea, D. and Dima, G.E., 2000. Anaphora and Cataphora: what's in there, to appear in *Proceedings of the 5th TELRI Seminar*, Ljubljana - Oct. 2000.

Cristea, D. and Dima, G.E., 2001. An Integrating Framework for Anaphora Resolution. *Information Science and Technology,* Romanian Academy Publishing House, Bucharest, vol. 4, no. 3.

Gordon, P.C. and Hendrick, R. 1997. Intuitive knowledge of linguistic coreference. *Cognition*, 62.

Gordon, P.C. and Hendrick, R. 1998. The Representation and Processing of Coreference in Discourse. *Cognitive Science,* 22.

Grosz, B.J., Joshi, A.K., and Weinstein, S. 1995. Centering: a Framework for Modelling the Local Coherence of Discourse, *Computational Linguistics,* 21 (2).

Hobbs, J.R. 1978. Resolving pronoun references. *Lingua*, 44. Also in B. Grosz, K. Sparck-Jones and B. Webber (eds.), *Readings in Natural Language Processing*, Morgan Kaufmann, Los Altos, 1986.

Keenan, E. and Comrie, B. 1977. Noun phrase accessibility and universal grammar. *Linguistic Inquiry*, 8.

Kennedy, C. and Boguraev, B. 1996. Anaphora for Everyone: Pronominal Anaphora Resolution without a Parser. *Proceedings of the 16th International Conference on Computational Linguistics*, vol.1.

Lappin, Y., Shalom, and Leass, Herbert J. 1994. An Algorithm for Pronominal Anaphora Resolution. *Computational Linguistics*, vol. 20, n. 4.

Mitkov, R.. 1997. Factors in Anaphora Resolution: They Are not the Only Things that Matter. A Case Study Based on Two Different Approaches. In R. Mitkov and B. Boguraev (eds.) *Proceedings of the Workshop "Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts"*, Universidad Nacional de Educación a Distancia, Madrid.

Mitkov, R, 2002. Anaphor resolution. Longman. Studies in Language and Linguistics.

Poesio, M., Vieira, R., and Teufel, S. 1997. Resolving bridging references in unrestricted texts. In R. Mitkov and B. Boguraev (eds.) *Proceedings of the Workshop "Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts"*, Universidad Nacional de Educación a Distancia, Madrid.

Reinhart, T. 1981. Definite NP anaphora and c-command domains. *Linguistic Inquiry*, 12.

Sidner, C. 1981. Focusing for interpretation of pronouns. *American Journal of Computational Linguistics*, 7.

Tufis, D. 2000. Using a Large Set of EAGLES-compliant Morpho-Syntactic Descriptors as a Tagset for Probabilistic Tagging. *Proceedings of LREC 2000*, Athens.

Vonk, W. 1985. The immediacy of inferences in the understanding of pronouns. In G. Rickheit and H. Stronher (eds.), *Inferences in text processing*.