About the effects of using Anaphora Resolution in assessing free-text student answers

Diana Pérez[1], **Oana Postolache**[3], **Enrique Alfonseca**[1],
Dan Cristea[2] and **Pilar Rodriguez**[1*]

[1]Dpt. of Computer Science, U. Autonoma Madrid (Spain)
[2]Dpt. of Computer Science, U. of Iasi (Romania)
[3]Dpt. of Computational Linguistics, U. of Saarland (Germany)

{Diana.Perez,Enrique.Alfonseca,Pilar.Rodriguez}@uam.es
dcristea@infoiasi.ro and oana@coli.uni-sb.de

Abstract

In this paper we present a possibility for integrating Anaphora Resolution (AR) in a system to automatically evaluate students' free-text answers. An initial discussion introduces some of the several methods that can be tried out. The implementation makes use of the AR-Engine RARE (Cristea *et al.* 02), integrated into the free-text answers assessor Atenea (Alfonseca & Pérez 04) to test these methods. RARE has been applied to find coreferential chains, and it has been found useful to extend the set of reference answers used by Atenea, by generating automatically new correct answers.

1 Introduction

Computer Assisted Assessment (CAA) is a field that studies how a computer can be used to assess students. One of its subfields, that has recently attracted much attention, focuses on assessing free-text answers. It is a quite complex task, still far from being completely solved. Thus, many systems are being developed, relying on various techniques. A classification of these techniques with examples of existing systems that use them is given in (Perez 04):

- Statistical techniques: they are based on some kind of statistical analysis, such as word frequency counts, or Latent Semantic Analysis (LSA) (Landauer *et al.* 01).
- Text Categorisation Techniques (TCT): they are applicable when the student's answer can be classified as right or wrong, or inside a category in a scale of grades, e.g. bad, intermediate, good and very good (Larkey 98).
- Information Extraction techniques: they are used by systems which acquire structured information from free text, for example dependencies between concepts as in Automark (Mitchell *et al.* 02).
- Full Natural Language Processing (NLP): NLP techniques, such as parsing and rhetorical analysis, can be used to gather more information about the student's answer. A system that applies NLP techniques is C-rater (Burstein *et al.* 01).
- Clustering: these techniques group essays that have similar words patterns to form a cluster with the same score. This is the approach followed by the Intelligent Essay Marking System (Ming *et al.* 00).
- Hybrid approaches: they combine several techniques to achieve better results. For instance, E-rater (Burstein *et al.* 98) and Atenea (Alfonseca & Pérez 04) use statistical and NLP techniques.

Although the techniques may seem very different, the general idea that underpins all these systems is the same: to compare the student's answer (or candidate answer) with the teacher's ideal answer (or reference answer). The closer they are, the higher the student's score is.

A problem to be able to compare the results of all these systems with each other is that, currently, there are not any standard evaluation corpora and metrics. Concerning the evaluation metrics, the one that is commonly used is the Pearson correlation between the teachers' and the system's scores on the same data set (Valenti *et al.* 03; Perez 04). The state-of-the-art results are between 30% and 93%, because the corpora used have very different degrees of difficulty.

Among the NLP techniques that can be employed to improve the automatic assessing of open-ended questions, Anaphora Resolution (AR), the process of finding the antecedent of an anaphora, could be considered as well. This language phenomenon, consisting of referring to a previously mentioned entity, is quite common in written language (Vicedo & Ferrández 00). Moreover, it has been successfully applied to other fields (Cristea *et al.* 05).

Previous authors have also mentioned that AR will probably be useful for free-text CAA (Valenti

[*] This work has been sponsored by Spanish Ministry of Science and Technology, project number TIN2004-0314.

et al. 03). However, to our knowledge, still there are no studies indicating the impact of applying AR to automatic assessment of free-text answers. Therefore, the main motivation of this paper is to study the effects of using AR integrated with the Atenea system. The AR-engine chosen is RARE (Cristea *et al.* 02). Our initial hypothesis was that somehow it would improve the accuracy of the assessment.

The first step to accomplish our aim has been to decide the way in which AR will be integrated with Atenea. The experimental framework given by the integration of RARE in Atenea has made possible to try several different uses of AR for free-text CAA. The indicator of the appropriateness of the procedure has been measured with the Pearson correlation between the teachers' and the system's scores. The results show that the application of AR directly on the student's answers does not improve the results in our case. On the other hand, AR has been found useful for generating automatically many alternative references and in this way, it slightly increases Atenea's assessment accuracy.

The paper is organised as follows: Section 2 presents the description of the possible uses that AR has in CAA of free-text answers. In Section 3, the implementation used in the experiments to test the previously mentioned methods is shown. Finally, in Section 4 several conclusions are drawn and future work is outlined.

2 Possible uses of AR in free-text CAA

Most of the systems for evaluating open-ended questions compare the student's candidate answer with reference answers written by the teachers. Therefore, the system will not be able to evaluate correctly an answer if the word choice or the expression used by the student and the teacher are different. We can try to solve this problem on both sides:

- Reducing the possible paraphrasings of each text, for instance, by eliminating all the pronouns and some definite NPs, using Anaphora Resolution.
- Extending the set of references with alternative paraphrasings. This can be done manually by asking several teachers to write alternative answers for the same question, or automatically, for instance, expanding the text with synonyms of the words used, or using AR, as described below.

Concerning the **reduction of paraphrasing** in a text, it is well known that there are many different expressions that have the same meaning. One of the sources for paraphrasing stems from the fact that there are many ways to refer to a previously mentioned entity by using an anaphoric expression. AR could help by identifying the referential expressions (REs) for the same referents, and gathering them in coreferential chains. Once coreferential chains are found, we have designed three ways in which they can be used:

1. *First-NP*: Each NP in the candidate and in the reference answers is substituted for the first NP in the coreferential chain. The aim is to filter the paraphrasing by substituting all NPs which refer to the same concept for the first NP used.

   For instance, let us suppose that we are scoring the candidate answer

   (1) Unix is an operating system. It is multiuser.

   and we apply this method to help in the comparison between this text and the references. The AR-engine RARE says that *Unix*, *operating system* and *It* are coreferential REs. Therefore, all of them will be substituted by the first RE (*Unix*). Therefore, the answer will be transformed into

   (2) Unix is Unix. Unix is multiuser

   Note that RARE considers the relationship between the subject and the predicative noun as coreferential as indicated in the MUC annotation guidelines (Hirschman *et al.* 97).

2. *All-NPs*: Each NP in the candidate and the reference answers is substituted for the whole coreferential chain to which it belongs. In this way, the candidate and reference answers will match if the intersection between the coreferential chains, considered as sets, is not empty. The third person singular personal pronouns *it* are excluded from these chains because most of the coreferential chains contain them.

   Thus, the candidate answer (1) will be transformed into

   (3) {an operating system,Unix} is {an operating system,Unix}. {an operating system,Unix} is multiuser

3. *Only-it*: Only the *it* pronouns in the candidate and the reference answers are substituted for the first NP in the coreferential chain which is not an *it*. This has been considered relevant enough to be studied given the extremely high frequency of this pronoun in the student answers in our test sets. This technique will also avoid the problem mentioned before with the predicative NPs.

   Thus, the resulting candidate answer for (1) would be

   (4) Unix is an operating system. Unix is multiuser.

Concerning the creation of new **reference answers** with alternative paraphrasings, we have also considered the possibility of applying AR in this task. While in the previous methods AR was applied to both the candidate and the reference answers, in this method it only affects the reference answers. The motivation is that the quality of the references is crucial, since they are the texts to which the students' answers are compared. Therefore, the usual practise of getting new references is to ask teachers to write these references.

However, as this is very cost and time consuming, we have also considered the automatic generation of new reference answers. It can be done by replacing automatically the NPs in the coreferential chains with other referential entities of those NPs. For instance, if we consider that (1) is a reference written by a teacher, two new references can be generated from its coreferential chain [Unix,an operating system,it]:"Unix is an operating system. Unix is multiuser" and "Unix is an operating system. An operating system is multiuser".

3 Implementation

3.1 Atenea

Atenea (Alfonseca & Pérez 04) is a CAA system for automatically scoring students' short answers. It has already been tested with English and Spanish texts and it could be easily ported to other languages. It works by processing the student's and teacher's answers according to several or all of the following NLP techniques, using the `wraetlic` tools (Alfonseca 03)[1]:

- **Stemming**: To be able to match inflected nouns or verbs.
- **Removal of closed-class words**: To be able to ignore them.
- **Word Sense Disambiguation**: To identify the sense intended by both the teacher and the student.

Then, the processed answers enter in the comparison module (ERB) that calculates the student's score and generates the student's feedback. This module is based on a modification of the n-gram co-occurrence scoring BLEU algorithm (Papineni *et al.* 01). The modification is necessary to take into account not only the precision but also the recall (Alfonseca & Pérez 04). The pseudocode of ERB is as follows:

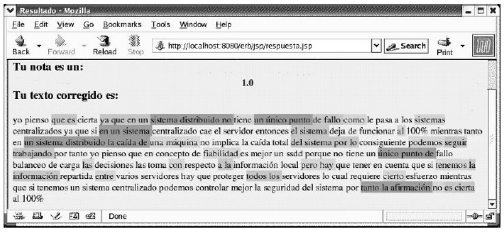
1. For each value of N (typically from 1 to 3), calculate the Modified Unified Precision (MUP_N) as the percentage of N-grams from the candidate answer that appears in any of the reference texts. It will be clipped by the maximum frequency with which it appears in any of the references.
2. Calculate the weighted linear average of MUP_N obtained for each value of N. Store it in $combMUP$.
3. Calculate the Modified Brevity Penalty (MBP) factor, which is intended to penalise answers with a very high precision, but which are too short, to measure the recall:
   (a) For N from a maximum value (e.g. 10) down to 1, look whether each N-gram from the candidate text appears in any reference. In that case, mark the words from the found N-gram, both in the candidate and in the reference.
   (b) For each reference text, count the number of words that are marked, and calculate the percentage of the reference that has been found in the student's answer.
   (c) The MBP factor is the sum of all those percentage values.
4. The final score is the result of multiplying the MBP factor by $e^{combMUP}$.

The answer will be returned to the student, together with a score and a feedback based on a colour code, in which the parts of the student's answer which appear in the references are marked with a darker background (see Figure 1).

3.2 RARE

RARE (Robust Anaphora Resolution Engine) allows the design, implementation and evaluation of different multilingual anaphora resolution models

[1] Available at www.ii.uam.es/~ealfon/eng/research/wraetlic.html


Resultado - Mozilla
File Edit View Go Bookmarks Tools Window Help
Back Forward Reload Stop
http://localhost:8080/erb/jsp/respuesta.jsp
Search
Print
Done

Tu nota es un:

1.0

Tu texto corregido es:

yo pienso que es cierta ya que en un sistema distribuido no tiene un único punto de fallo como le pasa a los sistemas centralizados ya que si en un sistema centralizado cae el servidor entonces el sistema deja de funcionar al 100% mientras tanto en un sistema distribuido la caída de una máquina no implica la caída total del sistema por lo consiguiente podemos seguir trabajando por tanto yo pienso que en concepto de fiabilidad es mejor un ssdd porque no tiene un único punto de fallo balanceo de carga las decisiones las toma con respecto a la información local pero hay que tener en cuenta que si tenemos la información repartida entre varios servidores hay que proteger todos los servidores lo cual requiere cierto esfuerzo mientras que si tenemos un sistema centralizado podemos controlar mejor la seguridad del sistema por tanto la afirmación no es cierta al 100%

Figure 1: Feedback for the student, and score.

on free texts. The engine (Cristea *et al.* 02; Postolache & Forascu 04) has successfully been integrated into a discourse parser (Cristea *et al.* 05) and a time tracking approach (Puscasu 04). It allows postponed resolution and deals with several varieties of anaphora from only pronominal anaphora to more complex types such as bridging anaphora. The information is organised in RARE on three layers:

1. **The text layer**: It is composed by the words that form the discourse and it is populated with the referential expressions (REs). For example, in the candidate answer "Unix is an operating system. It is multiuser", "Unix", "operating system" and "it" are the REs.
2. **The projection layer**: This layer stores information about the found REs in feature structures called projection structures (PSs) to help in determining which ones are coreferential.
3. **The semantic layer**: The REs represent entities from the real world. The underlying meaning of the REs is treated in the semantic layer on the form of Discourse Entities (DEs).

It is said that a PS is projected from an RE and a DE is proposed or evoked by a PS. The process should be done from left to right in languages that are read in that way and vice versa from those read from right to left. Irrespectively of the language, the necessary features for any AR model to be used in RARE are (Cristea & Dima 01):

- **A set of primary attributes**: indicating, for example, morphological, syntactic, semantic or positional information.
- **A set of knowledge resources**: such as a part-of-speech tagger and an NP extractor to fill in the primary attributes to be stored in the PSs.
- **A set of heuristics or rules**: for each RE they decide if it refers to a new DE or to an already existing one.

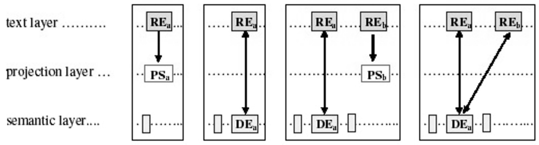


Figure 2: RARE layers.

- **A domain of referentiality**: it says where, how many and the order in which the DEs have to be checked.

The phases in the processing done by RARE are as follows (see Figure 2):

1. A referential expression RE_a is projected from the text layer into a feature structure PS_a on the projection layer. At this moment, the engine searches the space of existing discourse entities in order to recognise one against which the newly projected structure matches the best.
2. If no such DE is found, the projected structure PS_b is transformed in a new discourse entity DE_a, on the semantic layer, and disregarded from the projection layer. As the text unfolds, a new referential expression RE_b can be found on the text layer and, in its turn, projected as PS_b.
3. If PS_b matches an already existing discourse entity DE_a, with the meaning that their respective referential expressions, RE_a and RE_b, are coreferential. If this happens, PS_b is combined with DE_a and, subsequently, is disregarded from the projected layer.
4. Finally, chains of coreferential expressions are linked to the same object of the semantic layer, signifying that a unique discourse entity is evoked by all REs of the chain.

3.3 Techniques to use RARE in Atenea

The use of RARE as a new NLP module in Atenea requires the introduction of a new pre-initial phase to perform the pre-processing necessary to RARE. This phase includes a Functional Dependency Grammar (FDG) parsing of the text and the transformation of its result into an intermediate format understandable by RARE and Atenea. This format is a table in which each row represents a chain and, for each row, there are as many cells as NPs are in the chain. For the example candidate text (1) from Section 2, the equivalence table would have just one row (as it only has one chain) and it would be: *[Unix, an*

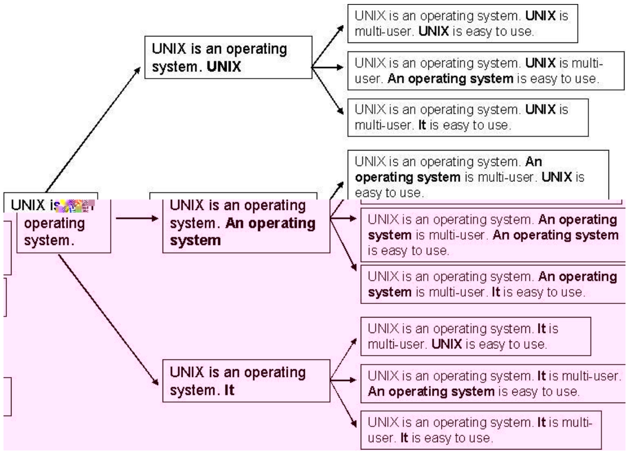

UNIX is an operating system.
UNIX is an operating system. UNIX
UNIX is an operating system. An operating system
UNIX is an operating system. It
UNIX is an operating system. UNIX is multi-user. UNIX is easy to use.
UNIX is an operating system. UNIX is multi-user. An operating system is easy to use.
UNIX is an operating system. UNIX is multi-user. It is easy to use.
UNIX is an operating system. An operating system is multi-user. UNIX is easy to use.
UNIX is an operating system. An operating system is multi-user. An operating system is easy to use.
UNIX is an operating system. An operating system is multi-user. It is easy to use.
UNIX is an operating system. It is multi-user. UNIX is easy to use.
UNIX is an operating system. It is multi-user. An operating system is easy to use.
UNIX is an operating system. It is multi-user. It is easy to use.

Figure 3: Example of the generation of new references from the original text "Unix is an operating system. It is multi-user. It is easy to use".

operating system, it].

The next step varies according to the method chosen. If it is is First-NP then each NP found in a row of the equivalence table is replaced by the first NP which is not an "it" in the chain. For All-NPs each NP found in a row of the equivalence table is replaced by the whole chain as a set. Finally, for Only-it each non-pleonastic "it" found in a row of the equivalence table is replaced by the first NP which is not an "it" in the coreferential chain.

Secondly, to implement the procedure for automatically generating *new paraphrases of the reference texts*, the following pseudocode has been used. It starts with one reference text that has been written by hand by a teacher.

```
1. Initialise an empty array genRefTexts with the reference text.
2. Look for the next non-pleonastic "it". If none is found, stop.
3. Identify the row of the table that contains the coreferential chain which includes the "it" pronoun found.
4. Create as many copies of all the references in genRefTexts as NPs exist in the coreferential chain. For each of the copies, the last "it" found has been replaced by each possible RE.
5. Go back to the second step.
```

Figure 3 shows an execution example.

3.4 Evaluation

For evaluation purposes, we have used a corpus composed of four sets of answers written by Spanish students in real exams about Operating Systems and a set of definitions of *Operating System*, retrieved from *Google glossary* in English.

N	NC	MC	NR	MR	Type
1	79	51	3	42	Def.
2	143	48	7	27	A/D
3	295	56	8	55	A/D
4	117	127	5	71	Y/N
5	38	67	4	130	Def.
M	134.4	69.4	5.4	65	-

Table 1: Answer sets used in the evaluation. Columns indicate: set number; number of candidate texts, mean length of the candidate texts (no. of words), number of references, mean length of the references, question type (Def.=definitions; A/D=advantages/disadvantages; Y/N=justified Yes/No).

N	ERB	S	C	S+C	W	W+C
1	0.5323	0.4337	0.5479	0.5310	0.4176	0.4841
2	0.6442	0.6899	0.6066	0.7567	0.6998	0.7655
3	0.2201	0.2426	0.3213	0.3459	0.2358	0.3282
4	0.3121	0.3326	0.3450	0.3754	0.3150	0.3586
5	0.5868	0.6007	0.5663	0.5702	0.6194	0.5919
M	0.4591	0.4599	0.4774	0.5158	0.4575	0.5057

Table 2: Results of Atenea without RARE, with (ERB) the statistical module, (S) stemming, (C), closed-class words removal, (W) word-sense disambiguation, and a combination of the previous procedures.

Because RARE only works in English, we have been forced to translate the first four datasets into English. The translation has been done with Altavista Babelfish[2]. In previous work (Pérez *et al.* 05), we have observed that the variation in accuracy of Atenea is not statistically significant when Babelfish is used to translate the texts (Pérez *et al.* 05), as the correlations got are very similar to the correlations when evaluating with the original texts.

The five data sets are described in Table 1. Table 2 shows the results, measured as the Pearson correlation between Atenea's scores and the teachers' scores, for several of Atenea's configurations without using RARE.

Reduction of paraphrasing The first experiment explores the impact of the reduction of paraphrasing both in the candidate answers and the references. The correlation between the teachers' and the system's scores has been calculated using the different settings of the system. The FDG-parsing of these data sets was done with the on-

[2] http://world.altavista.com/

N	FNP	ANP	It	NGR	ERB	S	C	S+C	W	W+C
1	0.5217	0.2506	0.5176	3	0.5212	**0.4688**	**0.5824**	**0.5501**	**0.4405**	**0.4951**
2	0.5984	0.5107	0.6337	8	0.6442	0.6355	**0.6667**	0.7094	0.6537	0.7199
3	0.1731	0.0209	0.1529	17	**0.2218**	0.2370	0.3083	0.3390	0.2255	0.3238
4	0.2102	0.1878	0.2222	13	0.2918	0.2853	**0.3806**	**0.4233**	0.2745	**0.4182**
5	0.5799	0.0239	**0.5941**	36	**0.5964**	**0.6141**	0.5607	**0.5903**	**0.6208**	**0.6054**
M	0.4167	0.1968	0.4241	15.4	0.4551	0.4481	**0.4997**	**0.5224**	0.443	**0.5125**
1m	**0.5806**	*0.4655*	**0.5498**	*3*	**0.5736**	**0.5373**	**0.5727**	**0.5597**	**0.5270**	**0.5608**

Table 3: Results achieved using Atenea with RARE. The first three columns show the results for reducing paraphrasing, using just the statistical ERB module: *First-NP* (FNP), *all-NPs*(ANP), and *only-It* (It). The other columns show the results when creating new references, tested with all of Atenea's configurations. Columns indicate the Number of Generated References (NGR), and the results with ERB, stemming (S), closed-class words removal (C), Word Sense Disambiguation (W) and several combinations between them. The last row, called *1m*, shows the results working with a manual translation of set 1 rather than with Babelfish's output

line demo of Connexor[3].

Table 3 (first three columns) shows the correlation values for different configurations of Atenea using RARE. The columns contain the results for each of the three heuristics. The bold font figure indicates the case in which using RARE has improved the result over the original ERB.

Contrary to our intuition, the results show that there is no significant improvement in using RARE and, in some cases, such as in the *all-NPs* method, the correlations decrease for all data sets. Therefore, our conclusion is that AR is not useful to improve the results of n-gram co-occurrence similarity metrics. However, as can be seen in row labelled *1m*, the correlations of all three strategies greatly improved when we work with a set of manual translations, and in two of them we obtain a higher correlation than when we worked without RARE (Table 2).

Creation of new references RARE has also been used to create new references by substituting the non-pleonastic it pronouns with all its Referential Expressions. Table 3, in its last seven columns, shows the results for several of Atenea's configurations. It can be seen that the use of RARE has improved three of the five configurations under test (C, S+C and W+C). Using RARE, the best configuration is to combine stemming and closed-class word removal.

Concerning the use of Set 1 translated manually, it can be seen that it also improves several of the configurations; however, the best results for set 1 are obtained when using the automatic translation and closed-class word removal, with a correlation that also exceeds that of the experiments on reduction of paraphrasing.

N	S	C	S+C	W	W+C
1	**0.4453**	**0.5677**	**0.4901**	**0.4195**	0.4356
2	0.6563	**0.6277**	0.6906	0.6756	0.7059
3	0.2288	0.2735	0.3192	0.2031	0.2746
4	**0.3449**	0.3126	0.3025	**0.3261**	0.2827
5	**0.6332**	0.5643	**0.5959**	**0.6529**	**0.6078**
M	**0.4617**	0.4692	0.4797	0.4554	0.4613

Table 4: Results achieved by Atenea using several NLP modules and the method of manually generating new references.

Finally, in order to study the effect of using RARE rather than any other Anaphora Resolution module, a last experiment has been performed by annotating the co-referential chains by hand. Table 4 shows that the results do not have a dramatic improvement, and even in some cases the correlation decrease when compared with the results using RARE. The reason is probably that RARE is probably more consistent in its answers (either correct or wrong) than a human annotator.

4 Conclusions and future work

In this paper, Anaphora Resolution has been applied to the task of automatically assessing students' free-text answers. In particular, the AR-engine RARE has been integrated into Atenea, to test four proposed methods: first-NP, in which the NPs are replaced by the first RE which is not the "it" pronoun; all-NPs, in which the NPs in the candidate and reference's texts are replaced by the whole coreferential chain; only-it, in which

[3] http://www.connexor.com/

only the "it" pronouns are replaced by the first RE; and the **automatic generation of variable references** from the original reference text, to automatically obtain new variants by replacing each non-pleonastic "it" with all the possible NPs in its coreferential chain.

From the results obtained, we can draw several interesting conclusions:

1. Previous findings indicated that BLEU-like algorithms produced consistent results on data that had been processed by MT engines (Pérez *et al.* 05). That was specially useful, in our case, in order to provide adaptation to the student's language without any intervention by the teacher. However, if we want to incorporate more sophisticated NLP steps, such as the reduction of redundancies using Anaphora Resolution, MT may not be adequate.

   On the other hand, MT is still acceptable using the procedure of automatic generation of references, in which the results increased with the use of RARE, and the best results have been obtained with the output of the MT engine.

2. The worst results have been obtained in the All-NPs configuration. We believe that it is due to the fact that the number of times that the candidate and the reference matches may be artificially inflated when the referential NPs are substituted by their REs. This is specially evident in the all-NPs experiment. We believe that there has not been much improvement because of the characteristics of the n-gram co-occurrence metric used.

3. Concerning the generation of new references, the results are slightly better, and the average correlation increases up to 52%. Furthermore, this method opens a promising line of future work that could be further exploited to automatically generate new references (for instance, with synonyms of the words in the references). Other lines of future work are the following: to improve the AR model with features specific to the types of answers to be processed, to finish the development of the Spanish anaphora resolution model for RARE, and to test more possibilities for using RARE with Atenea.

References

(Alfonseca & Pérez 04) E. Alfonseca and D. Pérez. Automatic assessment of short questions with a BLEU-inspired algorithm and shallow nlp. In *Advances in Natural Language Processing*, volume 3230 of *Lecture Notes in Computer Science*, pages 25–35. Springer Verlag, 2004.

(Alfonseca 03) E. Alfonseca. Wraetlic user guide version 1.0, 2003.

(Burstein *et al.* 98) J. Burstein, K. Kukich, S. Wolff, C. Lu, M. Chodorow, L. Bradenharder, and M. Dee Harris. Automated scoring using a hybrid feature identification technique. In *Proceedings of the Annual Meeting of the Association of Computational Linguistics*, 1998.

(Burstein *et al.* 01) J. Burstein, C. Leacock, and R. Swartz. Automated evaluation of essays and short answers. In *Proceedings of the International CAA Conference*, 2001.

(Cristea & Dima 01) D. Cristea and G.E. Dima. An integrating framework for anaphora resolution. *Information Science and Technology*, 4(3), 2001.

(Cristea *et al.* 02) D. Cristea, O. Postolache, G.E. Dima, and C. Barbu. Ar-engine - a framework for unrestricted co-reference resolution. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC)*, 2002.

(Cristea *et al.* 05) D. Cristea, O. Postolache, and I. Pistol. Summarisation through discourse parsing. In *Proceedings of CICLING 2005*, 2005.

(Hirschman *et al.* 97) Hirschman, Lynette, and Chinchor. Muc-7 coreference task definition, version 3.0. In *MUC-7 Proceedings*, 1997. See also: http://www.muc.saic.co.

(Landauer *et al.* 01) T.K. Landauer, D. Laham, and P.W. Foltz. The intelligent essay assesor: putting knowledge to the test. In *Proceedings of the Association of Test Publishers Computer-Based Testing: Emerging Technologies and Opportunities for Diverse Applications conference*, 2001.

(Larkey 98) L. S. Larkey. Automatic essay grading using text categorization techniques. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 90–95, 1998.

(Ming *et al.* 00) Y. Ming, A. Mikhailov, and T.L. Kuan. Intelligent essay marking system. *Learners Together*, 2000.

(Mitchell *et al.* 02) T. Mitchell, T. Russell, P. Broomhead, and N. Aldridge. Towards robust computerised marking of free-text responses, 2002.

(Papineni *et al.* 01) K. Papineni, S. Roukos, T. Ward, and W. Zhu. BLEU: a method for automatic evaluation of machine translation. Research report, IBM, 2001.

(Perez 04) D. Perez. Automatic evaluation of users' short essays by using statistical and shallow natural language processing techniques. Advanced Studies Diploma (Escuela Politécnica Superior, Universidad Autónoma de Madrid), 2004.

(Pérez *et al.* 05) D. Pérez, E. Alfonseca, and P. Rodríguez. Adapting the automatic assessment of free-text answers to the students profiles. In *Proceedings of the CAA conference*, Loughborough, U.K., 2005.

(Postolache & Forascu 04) O. Postolache and C. Forascu. A coreference model on excerpt from a novel. In *Proceeding of The European Summer School in Logic Language and Information - ESSLLI'2004*, Nancy, France, 2004.

(Puscasu 04) G. Puscasu. A framework for temporal resolution. In *Proceedings of the Language Resources and Evaluation Conference (LREC-2004)*, 2004.

(Valenti *et al.* 03) S. Valenti, F. Neri, and A. Cucchiarelli. An overview of current research on automated essay grading. *Journal of Information Technology Education*, 2:319–330, 2003.

(Vicedo & Ferrández 00) J.L. Vicedo and A. Ferrández. Importance of pronominal anaphora resolution to question answering systems. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 555–562, 2000.