Anaphora Resolution for Biomedical Literature by Exploiting Multiple Resources

Tyne Liang and Yu-Hsiang Lin

National Chiao Tung University, Department of Computer and Information Science, Hsinchu, Taiwan 300, ROC
{tliang, gis91534}@cis.nctu.edu.tw

Abstract. In this paper, a resolution system is presented to tackle nominal and pronominal anaphora in biomedical literature by using rich set of syntactic and semantic features. Unlike previous researches, the verification of semantic association between anaphors and their antecedents is facilitated by exploiting more outer resources, including UMLS, WordNet, GENIA Corpus 3.02p and PubMed. Moreover, the resolution is implemented with a genetic algorithm on its feature selection. Experimental results on different biomedical corpora showed that such approach could achieve promising results on resolving the two common types of anaphora.

1 Introduction

Correct identification of antecedents for an anaphor is essential in message understanding systems as well as knowledge acquisition systems. For example, efficient anaphora resolution is needed to enhance protein interaction extraction from biomedical literature by mining more protein entity instances which are represented with pronouns or general concepts.

In biomedical literature, pronominal and nominal anaphora are the two common types of anaphora. In past literature, different strategies to identify antecedents of an anaphor have been presented by using syntactic, semantic and pragmatic clues. For example, grammatical roles of noun phrases were used in [9] [10]. In addition to the syntactic information, statistical information like co-occurring patterns obtained from a corpus is employed during antecedent finding in [3]. However, a large corpus is needed for acquiring sufficient co-occurring patterns and for dealing with data sparseness.

On the other hand, outer resources, like WordNet[1], are applied in [4][12][15] and proved to be helpful to improve the system like the one described in [12] where animacy information is exploited by analyzing the hierarchical relation of nouns and verbs in the surrounding context learned from WordNet. Nevertheless, using WordNet alone for acquiring semantic information is not sufficient for solving unknown words. To tackle this problem, a richer resource, the Web, was exploited in [16]

[1] http://wordnet.princeton.edu/

R. Dale et al. (Eds.): IJCNLP 2005, LNAI 3651, pp. 742 – 753, 2005.

where anaphoric information is mined from Google search results at the expense of less precision.

The domain-specific ontologies like UMLS[2] (Unified Medical Language System) has been employed in [2] in such a way that frequent semantic types associated to agent (subject) and patient (object) role of subject-action or action-object patterns can be extracted. The result showed such kind of patterns could gain increase in both precision (76% to 80%) and recall (67% to 71%). On the other hand, Kim and Park [11] built their BioAR to relate protein names to SWISS-Prot entries by using the centering theory presented by [7] and salience measures by [2].

In this paper, a resolution system is presented for tackling both nominal anaphora and pronominal anaphora in biomedical literature by using various kinds of syntactic and semantic features. Unlike previous approaches, our verification of the semantic association between anaphors and their antecedents is facilitated with the help of both general domain and domain-specific resources. For example, the semantic type checking for resolving nominal anaphora can be done by the domain ontology UMLS and PubMed[3], the search engine for MEDLINE databases. Here, UMLS is used not only for tagging the semantic type for the noun phrase chunks if they are in UMLS, but also for generating the key lexicons for each type so that we can use them to tag those chunks if they are not in UMLS. If no type information can be obtained from an chunk, then its type finding will be implemented through the web mining of PubMed. On the other hand, the domain corpus, GENIA 3.02p corpus [20] is exploited while we solve the semantic type checking for pronominal anaphora. With simple weight calculation, the key SA/AO (subject-action or action-object) patterns for each type can be mined from the corpus and they turn out to be helpful in resolution. Beside the semantic type agreement, the implicit resemblance between an anaphor and its antecedents is another evidence useful for verifying the semantic association. Hence, the general domain thesaurus, WordNet, which supporting more relationship between concepts and subconcepts, is also employed to enhance the resemblance extraction.

The presented resolution system is constructed on a basis of a salience grading. In order to boost the system, we implemented a simple genetic algorithm on its selection of the rich feature set. The system was developed on the small evaluation corpus MedStract[4]. Nevertheless, we constructed a larger test corpus (denoted as '100-MEDLINE') so that more instances of anaphors can be resolved. Experimental results show that our resolution on MedStract can yield 92% and 78% F-Scores on resolving pronominal and nominal anaphora respectively. Promising results were also obtained on the larger corpus in terms of 87.43% and 80.61% F-scores on resloving pronominal and nominal anaphora respectively.

2 Anaphora Resolution

Figure 1 is the overview of the presented architecture, including the extraction of biomedical SA/AO patterns and semantic type lexicons in background processing

[2] http://www.nlm.nih.gov/research/umls/
[3] http://www.pubmedcentral.nih.gov/
[4] http://www.medstract.org/

(indicated with dotted lines), as well as the document processing, anaphor recognition and antecedent selection in foreground processing (indicated with solid lines).

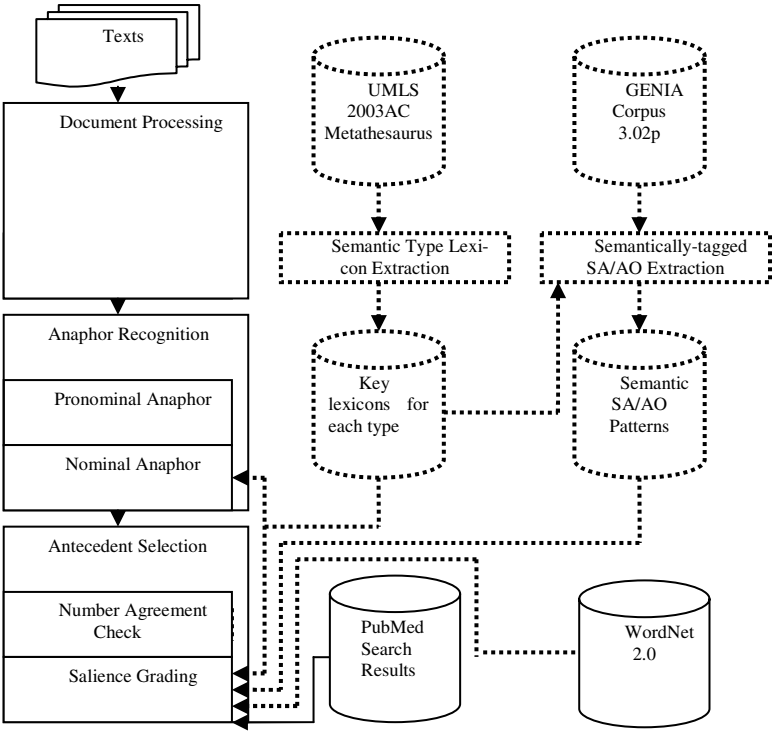

Fig. 1. System architecture overview

2.1 Syntactic Information Extraction

Being important features for anaphora resolution, syntactic information, like POS tags and base NP chunks, is extracted from each document by using the Tagger[5]. Meanwhile, each NP will be tagged with its grammatical role, namely, 'Oblique', 'Direct object', 'Indirect object', or 'Subject' by using the following rules which were adopted from [22] by adding rules 5 and 6.

Rule1: Prep NP (Oblique)
Rule2: Verb NP (Direct object)
Rule3: Verb [NP]$^{+}$ NP (Indirect object)
Rule4: NP (Subject) [",[^Verb], "|Prep NP]* Verb
Rule5: NP1 Conjunction NP2 (Role is same as NP1) Conjunction]
Rule6: [Conjunction] NP1 (Role is same as NP2) Conjunction NP2

[5] http://tamas.nlm.nih.gov/tagger.html

Rules 5 and 6 are presented for dealing those plural anaphors in such a way that the syntactic agreement between the first antecedent and its anaphora is used to find other antecedents. For example, without rules 5 and 6, 'anti-CD4 mAb' in Example 1 will not be found when resolving anaphora 'they'.

> Example1: "Whereas different anti-CD4 mAb or HIV-1 gp120 could all trigger activation of the ..., they differed…"

2.2 Semantic Information Extraction

Beside the syntactic clues, the semantic agreement between an anaphor and its antecedents can also facilitate anaphora resolution in domain-specific literature. In this paper, the semantic information for each target noun phrase chunk can be extracted with the help of the domain ontology, UMLS, which supports the semantic type for the chunk. However, the semantic types for those chunks which are not in UMLS are needed to be predicted. Therefore we need to extract the key lexicons from UMLS for each semantic type in background processing and use them to tag unknown chunk with predicted types. On the other hand, the semantic type checking for pronominal anaphors is done through the extraction of the key verbs for each semantic type. Hence, a domain corpus GENIA 3.02p is exploited in background processing.

2.2.1 Key Lexicons for Each Semantic Type

For each UMLS semantic type, its key lexicons are mined as the following steps in Figure 2:

A. Collect all UMLS concepts and their corresponding synonyms as type lexicon candidates.
B. Tokenize the candidates. For example, concept 'interleukin-2' has synonyms 'Costimulator', 'Co-Simulator', 'IL 2', and 'interleukine 2'. Then 'interleukin', 'costimulator', 'simulator', 'IL', and 'interleukine' will be treated as lexicon candidates.
C. For each candidate, calculate its weight w_{ij} for each type by using Eq. (1) which takes into account its concentration and distribution. A predefined threshold is given for the final selection of the candidates.

$$w_{i,j} = \frac{w_i}{Max\ c_j} \times \frac{1}{tw_i} \quad (1)$$

$w_{i,j}$: score of word i in semantic type j
w_i : count of word i in semantic type j
Max c_j : Max count of word k in semantic type j
tw_i : count of semantic types that word i occurs in

Fig. 2. Procedure to mine key lexicons for each semantic type

2.2.2 Semantic SA/AO Patterns

As indicated previously in Section 2.2, the semantic type checking for pronominal anaphors can be done through the extraction of the co-occurring SA/AO patterns extracted from GENIA 3.02p. We tagged each base noun phrase chunk from the corpus with its grammatical role and tagged it with UMLS-semantic type. Then we used Eq. 2 to score each pattern. At resolution, an antecedent candidate is concerned if its scores are greater than a given threshold. Table 1 is an example to show the key lexicons and verbs for two semantic types when the semantically-typed chunk is tagged with the role of subject.

$$score(type_i, verb_j) = \frac{frequency(type_i, verb_j)}{frequency(verb_j)} \times \frac{1}{No.\ of\ types(verb_j)} \quad (2)$$

Table 1. Some key lexicons and verbs for two semantic types

Semantic types	key lexicons for each type	key verbs for each type
Amino Acid, Peptide, or Protein	protein, product, cerevisiae, endonuclease, kinase, antigen, receptor, synthase, reductase, arabidopsis	bind, function, derive, raise, attenuate, abolish, present, signal, localize, release
Gene or Genome	gene, oncogenes	activate, compare, locate, regulate, remain, transcribe, encode, distribute, indicate, occupy

2.3 Anaphora Recognition

Anaphor recognition is to recognize the target anaphors by filtering strategies. Pronominal anaphora recognition is done by filtering pleonastic-it instances by using the set of hand-craft rules presented in [12]. On two corpora, namely, Medstract and the new 100-Medline corpus, 100% recognition accuracy was achieved. The remaining noun phrases indicated with 'it', 'its', 'itself', 'they', 'them', 'themselves' or 'their' are considered as pronominal anaphor. Others like 'which' and 'that' used in relative clauses are treated as pronominal anaphors and are resolved by the following rules.

Rule 1: 'that' is treated as pleonastic-that if it is paired with pleonastic-it.
Rule 2: For a relative clause with 'which' or 'that', the antecedents will be the noun phrases preceding to 'which' or 'that'.

On the other hand, noun phrases shown with 'either', 'this', 'both', 'these', 'the', and 'each' are considered as nominal anaphor candidates. Nominal anaphora recognition is approached by filtering those anaphor candidates, which have no referent antecedents or which have antecedents but not in the target biomedical semantic types. Following are two rules used to filter out those non-target nominal anaphors.

Rule 1: Filter out those anaphor candidates if they are not tagged with one of the target UMLS semantic types (the same types in [2])
Rule 2: Filter out 'this' or 'the' + proper nouns with capital letters or numbers.

We treated all other anaphors indicated with 'this' or 'the + singular-NP' as singular anaphors which have one antecedent only. Others are treated as plural nominal anaphors and their numbers of antecedents are shown in Table 2. At antecedent selection, we can discard those candidates whose numbers differ from the corresponding anaphors.

Table 2. Number of Antecedents

Anaphor	Antecedents #
Either	2
Both	2
Each	Many
They, Their, Them, Themselves	Many
The +Number+ noun	Number
Those +Number+ noun	Number
These +Number+ noun	Number

2.4 Antecedent Selection

2.4.1 Salience Grading

The antecedent selection is based on the salience grading as shown in Table 3 in which seven features, including syntactic and semantic information, are concerned.

Table 3. Salience grading for candidate antecedents

Features		Score
F1	recency 0, if in two sentences away from anaphor 1, if in one sentence away from anaphor 2, if in same sentence as anaphor	0-2
F2	Subject and Object Preference	1
F3	Grammatical function agreement	1
F4	Number Agreement	1
F5	Semantic Longest Common Subsequence	0 to 3
F6	Semantic Type Agreement	-1 to +2
F7	Biomedical antecedent preference	-2 if not or +2

The first feature *F1* is recency which measures the distance between an anaphor and candidate antecedents in number of sentences. From the statistics of the two corpora, most of antecedents and their corresponding anaphors are within in two sentence distance, so a window size for finding antecedent candidates is set to be two sentences in the proposed system. The second feature *F2* concerns the grammatical roles that an

anaphor plays in a sentence. Since many anaphors are subjects or objects so antecedents with such grammatical tags are preferred. Furthermore, the antecedent candidates will receive more scores if they have grammatical roles (feature *F3*) or number agreement (feature *F4*) with their anaphors.

On the other hand, features *5, 6,* and *7* are related to semantic association. Feature *5* concerns the fact that the anaphor and its antecedents are semantical variants of each other, so antecedents will receive different scores (as shown below) on the basis of their variation:

If there is total match of the semantic lexicons between an antecedent's head word and its anaphor
  Then salience score = salience score + 3
  Else If any antecedent component, other than head word, is matched with its anaphor
    Then salience score = salience score + 2
    Else If any antecedent component is matched with its anaphor's hyponym by WordNet 2.0
      Then salience score = salience score + 1

Following are examples to show the cases:

Example 2
case 1: total match:
  <anaphor: each *inhibitor*, antecedent: PAH alkyne metabolism-based *inhibitors*>
case 2: partial match:
  <Anaphor: both *receptor types*, antecedent: the ETB *receptor* antagonist BQ788>
case 3: component match by using WordNet 2.0:
  <Anaphor: this protein (hyponym: growth *factor*), antecedent: Cleavage and polyadenylation specificity *factor*>

If the antecedent can be found by UMLS,
  Then record its semantic types;
  Else If the antecedent contains the mined key lexicons of the anaphor's semantic type, then record the semantic type;
    Else mine the semantic type by web mining in such a way that searching PubMed by issuing {anaphor *Ana*, antecedent A_i } pair and applying Eq. 3 to grade its semantic agreement for A_i.

$$Score(A_i) = Score(A_i) - 1 + \left\lceil \frac{\#of\ pages\ containing(Ana, A_i)}{\#of\ pages\ containing(A_i)} \times 10 \right\rceil \times 0.3 \qquad (3)$$

Fig. 3. Procedure to find semantic types for antecedent candidates

Feature 6 is the semantic type agreement between anaphors and antecedents. As described in figure 3, the type finding for each antecedent can implemented with the help of UMLS. When there is no type information can be obtained from an antecedent, the type finding can be implemented with the help of PubMed, and the grading on such antecedent will be as Eq. 3. Feature 7 is biomedical antecedent preference. That is an antecedent which can be tagged with UMLS or the key lexicons database will receive more score.

2.4.2 Antecedent Selection Strategies

The noun phrases which precede a recognized anaphor in the range of two sentences will be treated as candidates and will be assigned with zero at initial state by the presented salience grader. Antecedents can be selected by the following strategies.

(1) Best First: select antecedents with the highest salience score that is greater than a threshold
(2) Nearest First: select the nearest antecedents whose salience value is greater than a given threshold

For plural anaphors, their antecedents are selected as follows:

(1) If the number of the antecedents is known, then select the same number of top-score antecedents.
(2) If the number of antecedents is unknown, then select those antecedent candidates whose scores are greater than a threshold and whose grammatical patterns are the same as the top-score candidate.

2.5 Experiments and Analysis

As mentioned in previous sections, a larger corpus was used for testing the proposed system. The corpus, denoted as '100-Medline', contains 100 MEDLINE abstracts including 43 abstracts (denoted as '43-Genia' in Table 6) randomly selected from GENIA 3.02p and another 57 abstracts (denoted as '57-PubMed' in Table 6) collected from the search results of PubMed (by issuing 'these proteins' and 'these receptors' in order to acquire more anaphor instances). There is no common abstract in the public MedStract and the new corpus. Table 4 shows the statistics of pronominal and nominal anaphors for each corpus.

Table 4. Statistics of anaphor and antecedent pairs

	Abstracts	Sentences	Pronominal instances	Nominal instances	Total
MedStract	32	268	26	47	73
43-GENIA	43	479	98	63	161
57-PubMed	57	565	69	118	187

The proposed approach was verified with experiments in two ways. One is to investigate the impact of the features which are concerned in the resolution. Another is to compare different resolution approaches. In order to boost our system, a simple

generic algorithm is implemented to yield the best set of features by choosing best parents to produce offspring.

In the initial state, we chose features (10 chromosomes), and chose crossover feature to produce offspring randomly. We calculated mutations for each feature in each chromosome, and evaluated chromosome with maximal F-Score. Top 10 chromosomes were chosen for next generation and the algorithm terminated if two contiguous generations did not increase the F-score. The time complexity associated with such approach is *O(MN)* where *M* is the number of candidate antecedents, *N* is number of anaphors.

Table 5. F-Score of Medstract and 100-Medlines

	Medstract						100-Medlines					
	Nominal			Pronominal			Nominal			Pronominal		
	P	R	F	P	R	F	P	R	F	P	R	F
Total Features	33/56	33/47		23/26	23/26		130/184	130/178		145/167	145/167	
	58.93	70.21	64.08	88.46	88.46	88.46	70.65	73.34	71.33	86.82	86.82	86.82
	F5, F6, F7			All-F5			F5, F6, F7			All-F5		
	P	R	F	P	R	F	P	R	F	P	R	F
Genetic Features	37/47	37/47		24/26	24/26		156/212	156/178		146/167	146/167	
	78.72	78.72	78.72	92.31	92.31	92.31	73.58	87.64	80.61	87.43	87.43	87.43

Table 6. Feature impact experiments

	Medstract		43-GENIA		57-PubMed	
	Nominal	Pronominal	Nominal	Pronominal	Nominal	Pronominal
All	64.08%	88.46%	67.69%	93.58%	73.28%	76.81%
All – F1	61.05%	73.08%	60.14%	83.87%	75.44%	75.36%
All – F2	65.96%	88.00%	70.22%	93.58%	78.40%	76.81%
All – F3	72.00%	80.77%	69.68%	84.46%	73.45%	76.81%
All – F4	64.65%	81.48%	68.33%	91.54%	73.73%	76.81%
All – F5	48.00%	92.31%	52.55%	93.58%	56.59%	78.26%
All – F6	44.04%	88.46%	46.42%	81.63%	57.14%	78.26%
All – F7	38.26%	59.26%	47.19%	71.96%	60.44%	50.72%

Table 5 shows that anaphora resolution implemented with the genetic algorithm indeed achieves higher F-scores than the one when all features are concerned. Table 5 also shows that the semantic features play more important role than the syntactic features for nominal anaphora resolution. Similar results can be also found in Table 6 where the impact of each feature is justified. Moreover, Table 6 indicates that the pronominal anaphora resolution on 43-Genia is better than that on the other two corpora. It implies that the mined SA/AO patterns from GENIA 3.02p corpus are

helpful for pronominal anaphora resolution. Moreover, Table 7 proves that the key lexicons mined from UMLS for semantic type finding indeed enhance anaphora resolution, yet a slight improvement is found with the usage of PubMed search results. One of the reasons is few unknown instances in our corpora.

On the other hand, comparisons with evaluation corpus, Medstract, were shown in Table 8 where the best-first strategy yielded higher F-score than the results by the nearest-first strategy. It also shows that the best-first strategy with the best selection by genetic approach achieves higher F-scores than the approach presented in [2].

Table 7. Impacts of the mined semantic lexicons and the use of PubMed

	With semantic lexicons		w/o semantic lexicons	
	Medstract.	100-Medlines	Medstract.	100-Medlines
With PubMed	78%	80.62%	59%	72.16%
Without PubMed	76%	80.13%	58%	71.33%

Table 8. Comparisons among different strategies on Medstract

	Best-First		Nearest-First		Castaño et al. [2]	
F-score	Nominal	Pronominal	Nominal	Pronominal	Nominal	Pronominal
Total Features	64.08%	88.46%	50.49%	73.47%		
Genetic Features	F5, F6, F7	All - F5	F5, F6, F7	All-(F2,F5)	F4, F5, F6	F4, F6, F7
	78.72%	92.31%	61.18%	79.17%	74.40%	75.23%

3 Conclusion

In this paper, the resolution for pronominal and nominal anaphora in biomedical literature is addressed. The resolution is constructed with a salience grading on various kinds of syntactic and semantic features. Unlike previous researches, we exploit more resources, including both domain-specific and general thesaurus and corpus, to verify the semantic association between anaphors and their antecedents. Experimental results on different corpora prove that the semantic features provided with the help of the outer resources indeed can enhance anaphora resolution. Compared to other approaches, the presented best-first strategy with the genetic-algorithm based feature selection can achieve the best resolution on the same evaluation corpus.

References

1. Baldwin, B.: CogNIAC: high precision coreference with limited knowledge and linguistic resources. In Proceedings of the ACL'97/EACL'97 workshop on Operational factors in practical, robust anaphora resolution (1997) 38-45
2. Castaño, J., Zhang J., Pustejovsky, H.: Anaphora Resolution in Biomedical Literature. In International Symposium on Reference Resolution (2002)

3. Dagan, I., Itai, A.: Automatic processing of large corpora for the resolution of anaphora references. In Proceedings of the 13th International Conference on Computational Linguistics (COLING'90) Vol. III (1990) 1-3
4. Denber, M.: Automatic resolution of anaphora in English. Technical report, Eastman Kodak Co. (1998)
5. Gaizauskas, R., Demetriou, G., Artymiuk, P.J., Willett, P.: Protein Structures and Information Extraction from Biological Texts: The PASTA System. Bioinformatics (2003)
6. Gasperin, C., Vieira R.: Using word similarity lists for resolving indirect anaphora. In ACL Workshop on Reference Resolution and its Applications, Barcelona (2004)
7. Grosz, B.J., Joshi, A.K., Weinstein, S.: Centering: A framework for modelling the local coherence of discourse. Computational Linguistics (1995) 203-225
8. Hahn, U., Romacker, M.: Creating Knowledge Repositories from Biomedical Reports:The MEDSYNDIKATE Text Mining System. In Pacific Symposium on Biocomputing (2002)
9. Hobbs, J.: Pronoun resolution, Research Report 76-1. Department of Computer Science, City College, City University of New York, August (1976)
10. Kennedy, C., Boguraev, B.: Anaphora for everyone: Pronominal anaphora resolution without a parser. In Proceedings of the 16th International Conference on Computational Linguistics (1996) 113-118
11. Kim, J., Jong, C.P.: BioAR: Anaphora Resolution for Relating Protein Names to Proteome Database Entries. ACL Workshop on Reference Resolution and its Applications Barcelona Spain (2004) 79-86
12. Liang, T., Wu, D.S.: Automatic Pronominal Anaphora Resolution in English Texts. Computational Linguistics and Chinese Language Processing Vol.9, No.1 (2004) 21-40
13. Mitkov, R.: Robust pronoun resolution with limited knowledge. In Proceedings of the 18th International Conference on Computational Linguistics (COLING'98)/ACL'98 Conference Montreal Canada (1998) 869-875
14. Mitkov, R.: Anaphora Resolution: The State of the Art. Working paper (Based on the COLING'98/ACL'98 tutorial on anaphora resolution) (1999)
15. Mitkov, R., Evans, R., Orasan, C.: A new fully automatic version of Mitkov's knowledge-poor pronoun resolution method. In Proceedings of CICLing- 2000 Mexico City Mexico (2002)
16. Modjeska, Natalia, Markert, K., Nissim, M.: Using the Web in Machine Learning for Other-Anaphora Resolution. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP2003) Sapporo Japan
17. Navarretta, C.: An Algorithm for Resolving Individual and Abstract Anaphora in Danish Texts and Dialogues. ACL Workshop on Reference Resolution and its Applications Barcelona, Spain (2004) 95-102
18. Ng, V., Cardie, C.: Improving Machine Learning Approaches to Coreference Resolution. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics (2002)
19. Oh, I.S., Lee, J.S., Moon, B.R.: Hybrid Genetic Algorithms for Feature Selection. IEEE Transactions on pattern analysis and machine Vol. 26. No. 11 (2004)
20. Ohta, T., Tateisi, Y., Kim, J.D., Lee, S.Z., Tsujii, J.: GENIA corpus: A Semantically Annotated Corpus in Molecular Biology Domain. In Proceedings of the ninth International Conference on Intelligent Systems for Molecular Biology (ISMB 2001) poster session (2001) 68

21. Pustejovsky, J., Rumshisky, A., Castaño, J.: Rerendering Semantic Ontologies: Automatic Extensions to UMLS through Corpus Analytics. LREC 2002 Workshop on Ontologies and Lexical Knowledge Bases (2002)
22. Siddharthan, A.: Resolving Pronouns Robustly: Plumbing the Depths of Shallowness. In Proceedings of the Workshop on Computational Treatments of Anaphora, 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003) (2003) 7-14
23. Yang, X., Su, J., Zhou, G., Tan, C.L.: Improving Pronoun Resolution by Incorporating Coreferential Information of Candidates. In Proceedings of ACL 2004 (2004) 127-134