Pronominal and Sortal Anaphora Resolution for Biomedical Literature

Yu-Hsiang Lin and Tyne Liang
Department of Computer and Information Science
National Chiao Tung University
Hsinchu, Taiwan
Email: gis91534@cis.nctu.edu.tw; tliang@cis.nctu.edu.tw;

Abstract. Anaphora resolution is one of essential tasks in message understanding. In this paper resolution for pronominal and sortal anaphora, which are common in biomedical texts, is addressed. The resolution was achieved by employing UMLS ontology and SA/AO (subject-action/action-object) patterns mined from biomedical corpus. On the other hand, sortal anaphora for unknown words was tackled by using the headword collected from UMLS and the patterns mined from PubMed. The final set of antecedents finding was decided with a salience grading mechanism, which was tuned by a genetic algorithm at its best-input feature selection. Compared to previous approach on the same MEDLINE abstracts, the presented resolution was promising for its 92% F-Score in pronominal anaphora and 78% F-Score in sortal anaphora.

1 Introduction

Anaphora resolution is one of essential tasks in message understanding as well as knowledge discovering. For example recognizing biomedical relations among biomedical entities from research literature like MEDLINE database requires anaphora resolution for those mentioned entities from texts.
There are different types of anaphora to be solved like pronominal, sortal (definite), zero, event, and coreference anaphora. In biomedical literature, pronominal anaphora and sortal anaphora are the two common anaphora phenomena. Pronominal anaphora is that mentioned entity is substituted by the pronoun. Sortal (definite) anaphora occurs in the situation that a noun phrase is referred by its general concept entity. Definite noun phrases are noun phrases stating with demonstrative articles, such as those, this, both, each and these or starting with a definite article.

Generally identifying antecedents of an anaphor can be handled by using syntactic, semantic or pragmatic clues. In past literature, syntax-oriented approaches for general texts can be found in [Hobbs, 76; Lappin and Leass 94; Kennedy and Boguraev 96] in which syntactic representations like grammatical role of noun phrases were used.

On the other hand more information other than syntactic information like co-occurring patterns obtained from the corpus was employed during antecedent finding in [Dagan and Itai, 90]. Information with limited knowledge and linguistic resources for resolving pronouns were found in [Baldwin, 97]. In [Denber, 98, Mitkov, 02], more knowledge from the outer resource like WordNet was employed in solving anaphora. Similarly WordNet together with additional heuristic rules were applied for resolving pronominal anaphora in [Liang and Wu, 04] which animacy information is obtained by analyzing the hierarchical relation of nouns and verbs in the surrounding context learned from WordNet.

In biomedical literature, it was found that sortal anaphors are prevalent in the texts like MEDLINE abstracts [Castaño et al., 02]. To deal this type of anaphora, Castaño et al. [02] used UMLS (Unified Medical Language System) as ontology to tag semantic type for each noun phrase and used some significant verbs in biomedical domain to extract most frequent semantic types associated to agent (subject) and patient (object) role of SA/AO-patterns. The result showed SA/AO-pattern could gain increase in both precision (76% to 80%) and recall (67% to 71%). In [Hahn et al., 02], a center list mechanism was presented to relate each noun to those nouns appearing in a previous sentence anaphora. Gaizauskas et al. [03] presented a predefined domain rules for ensuring co-referent between two bio-entities so that implicit relations between two entities could be recognized.

In this paper, the anaphora resolution for biomedical literature is achieved by employing UMLS ontology and syntactic information. The proposed system identifies both intra-sentential and inter-sentential antecedents of anaphors. In addition, anaphora resolution for unknown words has concerned in this paper by using headword mining and patterns mined from PubMed search results. Determining semantic coercion type of pronominal anaphor is done by SA/AO patterns, which were pre-collected from GENIA 3.02p corpus, a MEDLINE corpus annotated by Ohta et al. [02]. The final set of antecedents finding is decided with a salience grading mechanism, which is tuned by a genetic algorithm at its best-input feature selection. Compared to previous approach on the same MEDLINE abstracts, the presented resolution is promising for its 92% F-Score in pronominal anaphora and 78% F-Score in sortal anaphora.

2 The Presented Resolution

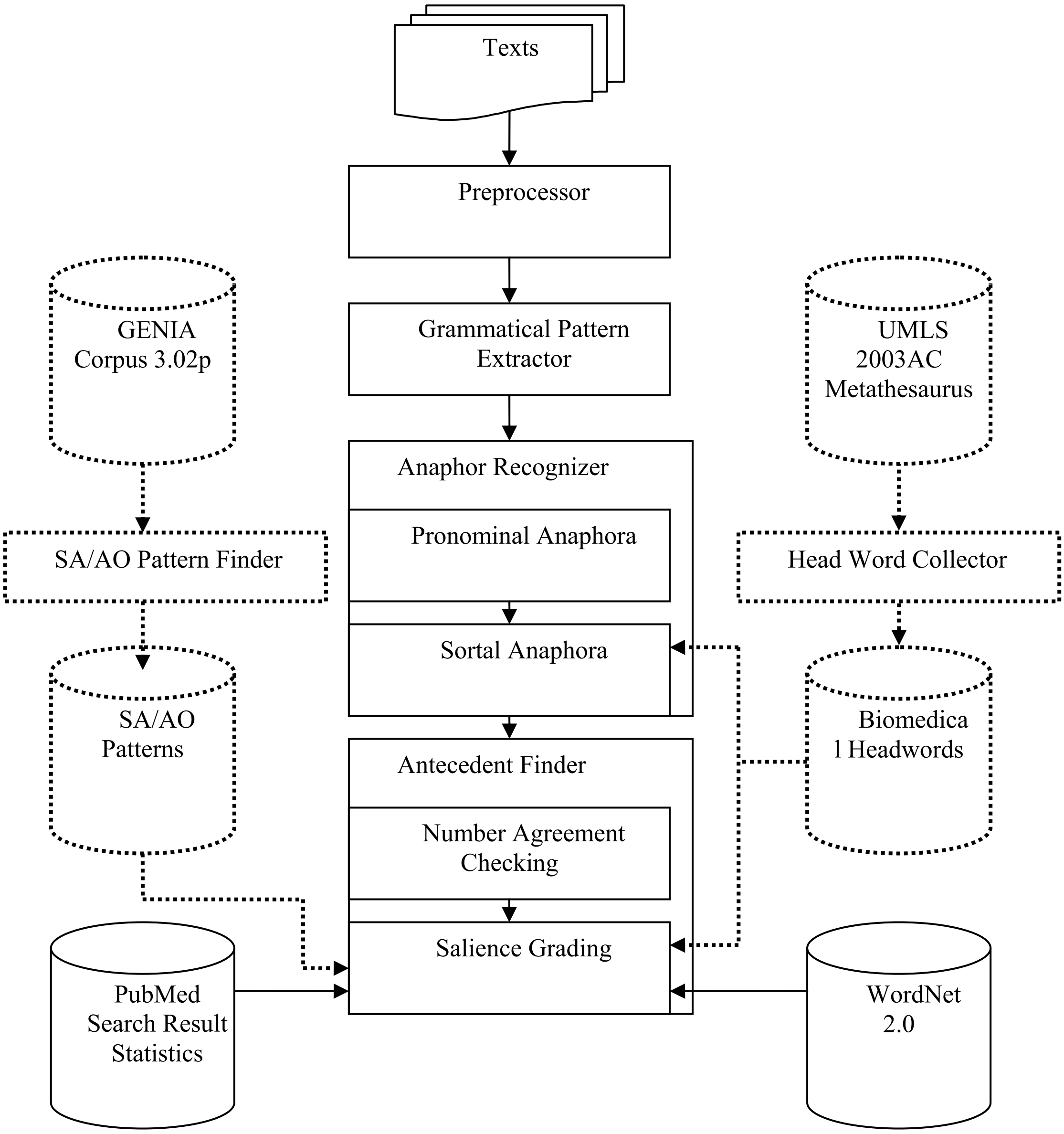

Figure 1: Architecture overview.

Figure 1 is the presented overview architecture which contains background processing, including SA/AO patterns and headword collection, indicated with dotted lines and foreground processing, including preprocessor, grammatical pattern extractor anaphor recognizer, and antecedent finder, indicated with solid lines.

2.1 SA/AO Patterns Collection

In this paper we used co-occurring SA/AO patterns obtained from GENIA corpus for pronominal anaphora resolution. Then we tag subjects and objects with UMLS-semantic type tags. Each SA/AO pattern is scored by the scoring function (Eq. 1). The antecedent candidates are concerned if their scores are greater than a given threshold.

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

The following is a pattern extraction example:

Example1:
<NFATp> <binds> to two sites within the kappa 3 element
UMLS semantic type of NFATp: Amino Acid, Peptide, or Protein
Extracted pattern: <Amino Acid, Peptide, or Protein> <bind>

2.2 Headword Collection

For unknown words, we need to predict their semantic types of the word. In [Pustejovsky et al., 02], they use the righthand head rule (the head of a morphologically complex word to be the righthand member of that word) to extract headwords to be subtype of the semantic type in UMLS. Table 1 is an example for headword 'receptor' which changes other noun phrase which were tagged with different semantic into 'Amino Acid, Peptide, or Protein'. 'Adhesion' is tagged with 'Acquired Abnormality, Disease or Syndrome' but 'adhesion receptor' becomes the tag of 'Amino Acid, Peptide, or Protein' by addition of 'receptor'.

Table 1: Example with righthand rule.

Noun Phrase	Semantic Type
Adhesion	Acquired Abnormality, Disease or Syndrome
adhesion receptor	Amino Acid, Peptide, or Protein
Contraction	Pathologic Function
Contraction receptor	Amino Acid, Peptide, or Protein
Estrogen	Steroid, Pharmacologic Substance, Hormone
estrogen receptor	Amino Acid, Peptide, or Protein
Dopamine	Organic Chemical…
dopamine receptor	Amino Acid, Peptide, or Protein

We collected all UMLS concepts and their corresponding synonyms, and then selected headwords for each semantic type (super-concept). For example, concept 'interleukin-2' has synonyms 'Costimulator', 'Co-Simulator', 'IL 2', and 'interleukine 2'. We collected 'interleukin', 'costimulator', 'simulator', 'IL', and 'interleukine' as headwords for 'interleukin-2'. Then, we found semantic types of 'interlukin-2' is 'Amino Acid, Peptide, or Protein' and 'Immunologic Factor'. We assigned synonym headwords of 'interleukin-2' into both semantic types. Eq. 2 was designed to score each headword for each semantic type. The scoring function smoothes the semantic type size.

Headword scoring function:

$$w_{i,j} = \frac{w_i}{Max \quad c_j} \times \frac{1}{tw_i} \quad (2)$$

$w_{i,j}$: 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

2.3 Preprocessor

After input untagged documents, we go through POS tagging and NP Chunking these preprocessing will give us more information about the documents.

2.4 Grammatical Function Extraction

Grammatical function is defined as creating a systematic link between the syntactic relation of arguments and their encoding in lexical structure. For anaphora resolution, grammatical function is an important feature of salience grading. We extended rules from Siddharthan [03], from following rules 1~4 to rules 1~6.

Rule 1: Prep NP (Oblique)
Rule 2: Verb NP (Direct object)
Rule 3: Verb [NP]$^+$ NP (Indirect object)
Rule 4: NP (Subject) [",[^Verb] appositive),"|Prep NP]* Verb
Rule 5: NP1 Conjunction NP2 (Role is same as NP1) Conjunction]
Rule 6: [Conjunction] NP1 (Role is same as NP2) Conjunction NP2

Rule 5 and rule 6 were presented for dealing those anaphors that have plural antecedents. We use syntactic agreement with first antecedent to find other antecedents. Without rules 5 and 6, 'anti-CD4 mAb' in Example 1 will not be found when resolving 'they''s antecedents.

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

3 Anaphora Resolution

Anaphor and antecedent recognition are the two main parts of the anaphora resolution system. Anaphor recognition is to recognize the target anaphora by filtering strategies. Antecedent recognition is to determine appropriate antecedents with respect to the target anaphor.

3.1 Anaphora Recognition

Noun phrases or prepositional phrases with 'it', 'its', 'itself', 'they', 'them', 'themselves' and 'their' are considered as pronominal anaphor. 'it', 'its', and 'itself' are considered as anaphor which has singular number of antecedent, others are considered as anaphor which has plural number of antecedents. Relative pronouns 'which' and 'that' are also pronominal anaphors but these anaphors can use a simple rule, point to the nearest noun phrase or prepositional phrase, to find its antecedent or point to the relative clause behind when paired with a pleonastic-it.

Noun phrases or prepositional phrases with 'either', 'this', 'both', 'these', 'the', and 'each' are considered as candidates of sortal anaphors. Noun phrases or prepositional phrases with 'this' or 'the+ singular noun' are considered as anaphors which have singular antecedent. Anaphor with plural number of antecedents are shown in Table 2.

Table 2: Number of Antecedents

Anaphor	Antecedents #
Either	2
Both	2
Each	Many
They, Their, Them, Themselves	Many
The +No.+ noun	No.
Those +No.+ noun	No.
these +No.+ noun	No.

3.1.1 Pronominal Anaphora Recognition

Pronominal anaphora recognition was done by filtering out pleonastic-it. Following rules are used to recognize pleonastic-it instances.

Rule1: It be [Adj|Adv| verb]* that

Example 2:
"It is shown that antibody 19 reacts with this polypeptide either bound to the ribosome or free in solution."

Rule 2: It be Adj [for NP] to VP

Example 3:
"However, it is possible for antidepressants to exert their effects on the fetus at other times during pregnancy as well as to infants during lactation."

Rule 3: It [seems|appears|means|follows] [that]*

Example 4:
"It seems that the presence of HNF1 sites in liver-specific genes was favoured, but that no counter-selection occurred within the rest of the genome."

Rule 4: NP [makes|finds|take] it [Adj]* [for NP]* [to VP|Ving]

Example 5:
"Furthermore, the same experimental model makes it possible to image lymphoid progenitors in fetal and adult hematopoietic tissues."

3.1.2 Sortal Anaphora Recognition

Sortal anaphora recognition was done by filtering those sortal anaphor, which have no referent antecedent or which have antecedents but not in the defined biomedical semantic types. Following two rules are used to filter out those un-target anaphors.

Rule 1: Filter out those noun phrases or prepositional phrases if they are not tagged with the following UMLS classes.

Amino Acid, Protein, Peptide, Embryonic Structure, Cell Biomedical Active Substance, Organism, Functional Chemical, Bacterium, Molecular Sequence, Chemical, Nucleoside, Cell Component, Enzyme, Gene or Genome, Structural Chemical Nucleotide Sequence, Substance, Organic Chemical, Pharmacologic Substance, Organism Attribute, Nucleic Acid, Nucleotide.

Rule 2: Filter out proper nouns with capitals and numerical features.

3.2 Number Agreement Checking

Number is the quantity that distinguishes between singular (one entity) and plural (numerous entities). It makes the process of deciding candidates easier since they must be consistent in number. All noun phrases and pronouns are annotated with number (singular or plural). For a specified pronoun, we can discard those noun phrases whose numbers differ from the pronoun. With singular antecedent anaphor, plural noun phrases are not considered as possible candidates.

3.3 Salience Grading

Salience grade for each candidate antecedent is assigned according to Table 3. Each candidate antecedent is assigned with zero at initial state.

Recency is a feature about distance between an anaphor and candidate antecedents. The closer between an anaphor and a candidate antecedent, the more chance the anaphor points to this candidate antecedent. For grammatical role agreement, if we use same entity in the second sentence and in the same role, it is easy for readers to identify which antecedent that the anaphor points to, so an author might use anaphor instead of full name of the entity. In addition to role agreement, subjects and objects are important role in sentence, which may be mentioned many times and writer might use an anaphor to replace a previously mentioned items. Singular anaphors may only point to one antecedent, while plural anaphors usually points to plural antecedents. For the feature of semantic type agreement, when we mention entity the second time, it is common for us to use its hypernym concept. Therefore such feature will receive high weights at salience grading.

Table 3: Salience grading for candidate antecedents.

Features	Score
Recency	0-2
Subject and Object Preference	1
Grammatical Role Agreement	1
Number Agreement	1
Longest Common Subsequence	0-3
Semantic Type Agreement	-1 if not or +2
Biomedical Antecedent	-2 if not or +2

3.3.1 Antecedent and Anaphor Semantic Type Agreement

For pronominal anaphora, we collected coercion semantic type between verb and headword by GENIA SA/AO patterns, and we generalized subjects and objects by using UMLS semantic types. For a pronoun, we tagged the pronoun with coercion semantic types on the basis of SA/AO pattern.

Sortal anaphoras are dealt by checking semantic agreement between anaphor and antecedent. So, all noun phrases and prepositional phrases will be tagged in advance by following steps.

(1) UMLS type check
(2) The Antecedent contains the headword in the anaphor's semantic type.
(3) If there is no headword found in antecedent then check {anaphor, antecedent} pair by using PubMed

For {anaphor, antecedent} pair {The nmd mouse mutation, of a second site suppressor allele}, we created query1 :<anaphor: "The nmd mouse mutation", antecedent: "of a second site suppressor allele"> and query2: <antecedent: "of a second site suppressor allele">. Queries are used to query from PubMed website and Eq. 3 was used to score the antecedent for semantic type agreement.

$$Score = -1 + \left[\frac{Query \quad pages \quad from \quad query\ 1}{Query \quad pages \quad from \quad query\ 2} \times 10 \right] \times 0.3 \tag{3}$$

3.3.2 Longest Common Subsequence (LCS)

The use of the LCS exploits the fact that the anaphor and its antecedents are morphological variants of each other (e.g., the anaphor "the grafts" and the antecedent "xenografts") [Castaño, 02]. We score each anaphor and candidate antecedent as follows:

```
If total match between a anaphor and its candidate antecedents
        then salience score = salience score + 3
Else if partial match between a anaphor and its candidate antecedents
        then salience score = salience score + 2
Else if one antecedent match its anaphor hyponym by WordNet 2.0
        then salience score = salience score + 1
```

3.3.3 Antecedent Selection

We search noun phrases or prepositional phrases in range of two sentences preceding the anaphor. We count salience grader scores for each noun phrase. Antecedents are selected by using best fit or nearest fit strategy.

(1) Best Fit: select antecedents with the highest salience score that is greater than threshold
(2) Nearest Fit: Select the nearest antecedents whose salience value is greater than a given threshold, and find candidate antecedents from the anaphor to the two sentences ahead

We have identified the number of antecedents for its corresponding anaphor. If an anaphor is identified to have plural antecedents, we will use following steps to choose antecedents.

(1) If the number of antecedents is identified, set the highest number of noun phrases or prepositional phrases to the anaphor.
(2) If the number of antecedents is unknown, find those noun phrases and prepositional phrases that are greater than a given threshold and they have the same patterns as the top-score noun phrase or prepositional phrase.

3.3.4 Feature Selection

Feature selection for salience grading was implemented with a genetic algorithm which can get the best features by choosing best parents to produce offspring leave local maximum by mutation.

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 found about two features to be mutated in each generation. Max F-Score was used to evaluate each chromosome and top 10 chromosomes were chosen for next generation. The algorithm terminated if two contiguous generations did not increase the F-score.

3.4 Experiments and Analysis

The test corpus, Medstract, was adopted from (http://www.medstract.org/), containing 32 MEDLINE abstracts and 83 anaphora pairs (26 pronominal and 57 sortal pairs). For pronominal anaphora, we tagged another 103 MEDLINE abstracts (103-MEDEDLINSs) corpus which contains 177 pronominal anaphora pairs.

From the experimental results in Table 4, best fit strategy performed better than the nearest first strategy. In addition, the features selected by the genetic algorithm indicated that syntactic features affect pronominal anaphora, and semantic features will impacts on both sortal and pronominal anaphora.

Table 4: System result with best-first and nearest-first algorithm for Medstract.

	Best Fit		Nearest Fit		[Castano et al., 2002]	
	Sortal	Pronominal	Sortal	Pronominal	Sortal	Pronominal
Total Features	64.08%	88.46%	50.49%	73.47%		
Genetic Features	F5~F7	All-{F5}	F5~F7	All-{F2,F5}	F4~F6	F4, F6, F7
	78.26%	92.31%	61.18%	79.17%	74.4%	75.23%

F1: Recency, F2: Subject and Object preference, F3: Grammatical role Agreement, F4: Number Agreement, F5: Longest common subsequence, F6: Semantic type Agreement, F7: Biomedical Antecedent

The impact of each feature was also concerned and verified with the same corpus. Syntactic features (F1~F4) play insignificant roles in sortal resolution but they are useful for pronominal anaphora resolution. Sortal anaphora resolution are sensitive to semantic features (F5~F7), semantic type agreement plays an important role in sortal anaphora resolution. In addition to UMLS, headwords and PubMed search results were used to determine semantic type agreement between anaphor and antecedents. Table 5 shows F3 increases F-score in pronominal anaphora but drop F-score in sortal anaphora. Medstract and 103-MEDLINEs results show semantic type match is important in both sortal and pronominal anaphora. Table 6 shows F-score when removing headword and PubMed query result. Headword features show improvement in F-score because the semantic type of new words become precisely. PubMed query results improved little in F-score may because we only use co-occurrence information was concerned.

Table 5: Impact of each feature in pronominal and sortal.

	Medstract		103-MEDLINEs
	Sortal	Pronominal	Pronominal
All	64.08%	88.46%	85.88%
All – Recency (F1)	61.05%	73.08%	79.10%
All - Subject or Object preference (F2)	65.96%	88.00%	84.18%
All - Grammatical Role Match (F3)	72.00%	80.77%	80.79%
All - Number Agreement (F4)	64.65%	81.48%	85.88%
All – LCS (F5)	48.00%	92.31%	86.44%
All – Semantic Type Match (F6)	44.04%	88.46%	77.40%
All - Biomedical Antecedent (F7)	38.26%	59.26%	61.02%

Table 6: Impact of headword and PubMed.

	With Headword	Without Headword
With PubMed	78%	59%
Without PubMed	76%	58%

4 Conclusion

In this paper, pronominal and sortal anaphora which are common phenomenal in biomedical texts are discussed. The pronominal anaphora processing was achieved by syntactic and semantic features, while sortal anaphora was tackled by semantic features. For new biomedical entities to UMLS, we solve the entities semantic agreement by using headword mining and patterns mine from PubMed query results. Experiment results showed the proposed strategies indeed enhance the resolution in terms of higher F-Score.

Acknowledgement

This research is partially supported by MediaTek Research Center, National Chiao Tung University, Taiwan.

5 References

[1] Breck Baldwin, “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, pp. 38-45.

[2] José Castaño, Jason Zhang, Hames Pustejovsky, “Anaphora Resoution in Biomedical Literature,” *In International Symposium on Reference Resolution,* 2002

[3] Ido Dagan and Alon Itai, “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, 1-3,* 1990.

[4] Michel Denber, “Automatic resolution of anaphora in English,” *Technical report, Eastman Kodak Co.* , 1998.

[5] Udo Hahn and Martin Romacker, “Creating Knowledge Repositories from Biomedical Reports:The MEDSYNDIKATE Text Mining System, “*In Pacific Symposium on Biocomputing, 2002*

[6] J. Hobbs, “Pronoun resolution,” *Research Report 76-1, Department of Computer Science, City College, City University of New York, August 1976*

[7] R. Gaizauskas, G. Demetriou, P.J. Artymiuk and P. Willett, “Protein Structures and Information Extraction from Biological Texts: The PASTA System,” *In Bioinformatics 2003*

[8] Christopher Kennedy and Branimir Boguraev, “Anaphora for everyone: Pronominal anaphora resolution without a parser,” *In Proceedings of the 16th International Conference on Computational Linguistics*, 1996, pp.113-118.

[9] Shalom Lappin and Herbert Leass, “An Algorithm for Pronominal Anaphora Resolution,” *Computational Linguistics*, Volume 20, Part 4, 1994, pp. 535-561.

[10] Tyne Liang and Dian-Song Wu, “Automatic Pronominal Anaphora Resolution in English Texts,” *In Computational Linguistics and Chinese Language Processing Vol.9, No.1, 2004, pp. 21-40*

[11] Ruslan Mitkov, “Robust pronoun resolution with limited knowledge, “ *In Proceedings of the 18th International Conference on Computational Linguistics (COLING'98)/ACL'98 Conference Montreal, Canada.* 1998, pp. 869-875.

[12] Ruslan Mitkov, “Anaphora Resolution: The State of the Art,” *Working paper (Based on the COLING'98/ACL'98 tutorial on anaphora resolution)*, 1999.

[13] Ruslan Mitkov and Catalina Barbu, “Evaluation tool for rule-based anaphora resolution methods,” *In Proeedings of ACL'01, Toulouse*, 2001.

[14] Ruslan Mitkov, Richard Evans and Constantin Orasan, “A new fully automatic version of Mitkov's knowledge-poor pronoun resolution method,” *In Proceedings of CICLing- 2000, Mexico City, Mexico*.

[15] T. Ohta, Y. Tateisi, J.D. Kim, S.Z. Lee and J. Tsujii. “GENIA corpus: A Semantically Annotated Corpus in Molecular Biology Domain.” *In the Proceedings of the ninth International Conference on Intelligent Systems for Molecular Biology (ISMB 2001) poster session. pp. 68. 2001.*

[16] James Pustejovsky, Anna Rumshisky, José Castaño, ” Rerendering Semantic Ontologies: Automatic Extensions to UMLS through Corpus Analytics,” *LREC 2002 Workshop on Ontologies and Lexical Knowledge Bases,* 2002.

[17] J. Pustejovsky, José Castaño, J. Zhang, B. Cochran, M. Kotecki, ” Robust Relational Parsing over Biomedical Literature: Extracting Inhibit Relations.,” *In Pacific Symposium on Biocomputing, 2002*