生物文獻中同指涉問題處理之研究

研究生：林裕祥　　指導教授：梁婷 博士

國立交通大學資訊科學系

摘要

同指涉消解需要處理指代現象消解和縮寫鏈結串聯。我們使用規則式處理縮寫問題，這規則式處理法則包含七條規則和使用了名詞片語辨識器(NP-chunker)來辨識縮寫和縮寫的原型。我們可以處理縮寫問題達到 97%正確率和 88%的招回率。除了縮寫問題，我們處理了在生物文獻中常見的代名詞指代和名詞指代詞問題。處理機制裡加入了知識本體(UMLS)和從生物文獻中探勘出來的 SA/AO (subject-action/action-object)樣板。在此同時，對於名詞指代現象中未知詞使用了從 UMLS 中收集的中心詞(headword)和從 PubMed 中探勘的樣板。我們用基因演算法所得出了最佳特徵值給分機制，來決定指代詞和和它先行詞的關係。與其它方法在相同語料(MEDLINE 摘要)做比較，所提的方法處理指代詞指代現象可達到 92% F-Scorec 和名詞指代現象可達到 78% F-Score。

Coreference Resolution in Biomedical Literature

Student：Yu-Hsiang Lin Advisor：Tyne Liang

Department of Computer and Information Science

National Chiao Tung University

Abstract

Coreference resolution involves anaphora and abbreviation linkage. To handle abbreviations, we use a rule-based resolution which concerns seven rules with the help of a NP-chunker to identify abbreviation and its long form. Our abbreviation resolution can achieve 97% in precision and 88% in recall. On the other hand, we address pronominal and sortal anaphora, which are common in biomedical texts. The resolution was achieved by employing the 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 stage. Compared to previous approaches 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.

ACKNOWLEDDGEMENTS

Thank my advisor Dr. Tyne Liang for her encouragement and teaching during the two and half years, and then this thesis can be accomplished. Besides, she gives me a guide to make plans for the future.

Also, I thank the members of information retrieval laboratory including Chien-Pang Wang, Dian-Song Wu, Chih-Chien Chao, Ping-Ke Shih , Lan-Chi Lin, Yi-Li Chen, Yi-Chia Wang, and Hsiao-Ju Shih. They provide suggestions and comments to assist in writing and experimenting.

Finally, I thank my family for support and encouragement in the period of school life.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1.

Introduction

1.1. Motivation and Goal

Coreference resolution is one of essential tasks in message understanding. We introduce a method for adding abbreviation to anaphora resolution system to create coreference link. In this thesis, we deal with abbreviation and anaphora for biomedical literature. Unlike previous approach [Yoshida et al. 00] which used a NE tagger with six rules, our approach uses a rule-based resolution which concerns seven rules with the help of a NP-chunker to identify abbreviation and long form pair. Our abbreviation resolution can achieved 97% in precision and 88% in recall.

On the other hand, anaphora resolution is achieved by employing UMLS ontology and syntactic information. Our anaphora resolution approach identifies both intra-sentential and inter-sentential antecedents of anaphors. In addition, anaphora resolution for unknown words has concerned in this thesis 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. [01]. 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 Castaño et al. [02] 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.

1.2. Problem Definition

According to the Coreference Task Definition of MUC-6 and MUC-7 [1] the coreferences may be in one of the following types:

1) Basic Coreference: The basic criterion for linking two markables is whether they are coreferential: whether they refer to the same object, set, activity.

Example 1: <p70 S6 kinase> <(p70(S6k))> is a mitogen-activated protein kinase… a neuronally enriched protein of 1,095 amino acids that contains a PDZ domain and binds <p70(S6k)>

2) Bound Anaphors: MUC-6 and MUC-7 also make a coreference link between a "bound anaphor" and the noun phrase which binds it (even though one may argue that such elements are not coreferential in the usual sense).

Example 2: <every man> who knows <his> own mind.

3) Apposition: A typical use of an appositional phrase is to provide an alternative description or name for an object:

Example 3: “Julius Caesar”, “the well-known emperor”.

4) Predicate Nominatives and Time-dependent Identity: Predicate nominatives are

[1] http://www.itl.nist.gov/ia[1]ui/894.02/related_projects/muc/proceedings/co_task.html

also typically coreferential with the subject.

Example 4: <Bill Clinton> is <the President of the United States>.

On the other hand, Kulick et al., [03] defined coreference as the following types:

1) Anaphor: Pronouns or definite (sortal) NPs used as anaphors.
2) Is-a Relation: This includes cases of predicate nominal and appositives, such as C-kit, a tyrosine kinase…. By separating this out from "anaphor", they maintain the constraint that members of a coreference (anaphor or acronym) chain must be in an equivalence relation.
3) Acronym Definition: The usage of an acronym points back to the antecedent where it is defined.
4) Acronym Linkage: Acronyms are linked together, with the first occurrence in turn pointing to the definition of the acronym with an acronym definition link.

It is noticed that the definitions 1-4 given by Kulick et al., [03] are relevant to the definition 1 and 3 given by MUC-6 and MUC-7.

Coreference needs to solve anaphora and term variants problems. In [Jacquemin C. and Tzoukermann E., 97], term variants can be abbreviation (e.g. 'p70(S6k)' and 'p70 S6 kinase') , permutations variant (e.g. 'protein of muscle' and 'muscle protein'), and coordination variant (e.g. 'human chromosomes 11p15 and 11p13' and 'human chromosomes 11p15 and human chromosomes 11p13'). In this thesis, we deal with abbreviation, pronominal anaphora and sortal anaphora.

There are different types of anaphora to be solved like pronominal, sortal (definite), zero, and event anaphora. In biomedical literature, pronominal anaphora and sortal anaphora are the two common anaphora. From the study on ten Medline documents, we found that there are about 52% pronominal and 46% sortal anaphors. Pronominal anaphora is mentioned entity which is substituted by a pronoun.

This type of anaphora can be divided into following subclasses:

Nominative: {he, she, it, they}

Reflexive: {himself, herself, itself, themselves}

Possessive: {his, her, its, their}

Objective: {him, her, it, them}

Relative: {who, whom, which, that}

Example 5: <CsA> and <FK506> are powerful suppressors of the immune system, …

<They> act at a point in activation that lies between receptor ligation …

Sortal (definite) anaphora occurs in situation that a noun phrase is referred by its general concept entity. The general concept entity can be a semantically close phrase such as synonyms or superordinates of the antecedent [Mitkov, 99]. Definite noun phrases are noun phrases starting with demonstrative articles, such as 'those', 'this', 'both', 'each', 'these' and 'the'. Example 6 is an example of sortal anaphora:

Example 6: <The human transcription unit> and <mouse GATA3 transcription unit> start at a major initiation site. The promoter sequence analysis of <these two genes> revealed…

1.3. Previous Works

Cardie and Wagstaff [1999] treated coreference resolution as a clustering task. This approach used a feature-based distance function to verify any pair of NPs to be a coreference or not.

In the recent literatures [Aone and Bennett, 95; McCarthy and Lehnert, 95; Soon

et al., 01; Ng and Cardie, 02a; Ng and Cardie, 02b; Ng and Cardie, 02c, Ng 04], coreference resolution is framed as classification. Decision tree algorithms C4.5 and RIPPER were applied in [Ng and Cardie, 02a] on the two standard coreference resolution data sets, MUC-6 and MUC-7, with F-Score of 70.4% and 64.3%. In [Ng and Cardie, 02b] and [Ng, 04], C4.5 was used to identify anaphoric and non-anaphoric noun phrases, and the result shows that anaphoricity information can improve the precision at the expense of lower recall. The standard coreference (MUC-6 and MUC-7) data sets contain only about 5% positive instances. Ng and Cardie [02c] combine negative sample selection, positive sample selection and error-driven pruning for machine learning of coreference rules. It turns out to improve F-Scores from 52.4% to 69.5% for MUC-6 corpus and 41.3% to 63.4% for MUC-7 corpus.

To deal with term variants, rule-base methods were presented to resolve coordination variants in GENIA corpus [Shih, 2004]. For abbreviation resolution, Yoshida [et al., 00], they used a rule based protein name tagger (PROPER) and six types to extract protein names and their abbreviations. Pustejovsky et al. [01] presented NP chunk-based identification for the boundary of a long form. When a noun phrase was found to precede an abbreviation, each of the characters within the abbreviation was matched in the long form. A grading function is used to take into account. The match is accepted if the score is greater than a given threshold. Schwartz and Hearst [03] used ‘long form (short form)’, and ‘short form (long form)’ to identify candidate abbreviation pairs. The abbreviation recognition is starting from the end of the short form and the long form, move right to left, trying finding the shortest long form that matches the short form. Every character in the short form must match a character in the long form, and the matched characters in the long form must be in the same order as the characters in the short form.

After recognition abbreviations, we need to solve anaphora phenomena to create coreference chain. Anaphora resolution is to identify antecedents of an anaphor. It 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 information 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 [Markert et al., 03], instead of using handcrafted lexical resources, they search the Google with shallow pattern which can be predetermined for the type of anaphoric phenomenon. Coreference information can be used to identify the type of an anaphor. Yang et al. [04] added information of coreferent NP as features to select antecedents of anaphor.

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.

Chapter 2.

Abbreviation Resolution

An abbreviation is a letter or group of letters, taken from a word or words, and employed to represent them for the sake of brevity. Acronym can be treated as a special case of abbreviation.

2.1. NP Chunking

In this thesis, we use the English Part-of-Speech Tagger (http://tamas.nlm.nih.gov/tagger.html) proposed by Tamas Doszkocs, Ph.D. For abbreviation resolution, we use the long definition of noun phrase such as 'NP [PP*]' to be one NP.

2.2. Abbreviation Candidates Identification

After NP chunking, we use syntactic constraints presented by Pustejovsky et al. [01] to find abbreviation pairs.

1) NP (NP)

   nitroglycerin (NTG)

2) NP, NP.

   nitroglycerin, NTG.

3) NP,NP, w/o conjunction

   nitroglycerin, NTG,

2.3. Long Form Chunking

Since abbreviations have fewer characters than their long forms, the shorter NP is considered as the abbreviation and the other NP is considered as candidate long form of the abbreviation. To chunk the best long form, we use six rules from Yoshida [et al., 00] and added 'Rule 7' as list below.

Rule 1: A long form consists of initial characters.

Thyrotrophin-releasing hormone (TRH)

Rule 2: A long form consists of capital or numerical characters.

IL2 receptor sub-unit (IL-2R)

Rule 3: A long form consists of initial characters of syllables.

nitroglycerin (NTG)

Rule 4: A long form consists of some characters of which an abbreviation is composed

Megestrol acetate (megace)

Rule 5: The order of some characters in an abbreviation is changed (inversion)

NG-monomethyl-L-arginine (L-NMMA)

Rule 6: Some characters in abbreviation are substituted by other expressions in long form.

fibronectin type III (FN3)

Rule 7: A long form consists of end characters.

candidate boundary elements (cBEs)

After identifying seven rules, we use the long form chunking steps as show in Figure 1. For each candidate abbreviation and long form pair, the identification time complexity is O(nM), where n is number characters of candidate abbreviation, and M is number of syllables in candidate long form. For each tokens we need go through all

seven rules and all rules are equal weighted.

```
Abbreviation_LongForm_Cunking(NP1, NP2)
    Input: NP1, NP2
    Output: NP1 (abbreviation), Word (Long form)
    Method:
    NO_NP1_Chars← Number of characters of NP1
    NO_NP2_Chars←Number of characters of NP2
    //ensure NP1 be the candidate
    if NO_NP1_Chars > NO_NP2_Chars
      switch NP1 and NP2, switch NO_NP1_Chars and NO_NP2_Chars
    end if
    m← 0
    Long←Words of NP2
    W←Number of Words of NP2
    Abbreviation←Characters of NP1
    while ( m small or equal to W )       //Chunk words from right to left word
          Word←Long[W-m] to Long[W]
          NO← Using rule 1~7 to check number of characters of Abbreviation
              can fit in Word
          if( NO/ NO_NP1_Chars > threshold )
              break;
          end if
          m++
    end while
if the initial character of Long not equal initial character of Abbreviation
          if initial character of Long[W-m-1] eq initial character ofAbbreviation
              Word←Long[W-m-1] to Long[W]
          end if
end if
return NP1, Word
```

Figure 1: Algorithm identifying abbreviation and chunking long form.

2.4. Experimental Results and Analysis

We tested our abbreviation recognition system by two corpora, the Medstract Gold Standard Evaluation Corpus (Medstract), which has 133 abstracts including 168 <abbreviation form, long form> pairs, and 100-MEDLINS which includes 162 <abbreviation form, long form> pairs. Information of the two corpora are shown in Table 1. From Table 1 we show the information of Medstract and 100-Medlines. It shows we have 298 'NP(NP)' and 143 'NP, NP' candidates in Medstract. For 100-Medlines we have 380 'NP(NP) and 'NP, NP' candidates. Amount these candidates, we have 163 'LF(Ab)', 2 'Ab(LF)', and 3 'NP, NP' pairs are truly abbreviation and long forms pairs in Medstract. In 100-Medlines we have 162 'LF(Ab)', 0 'Ab(LF)', and 0 'NP, NP' pairs are truly abbreviation and long form pairs.

Table 1: Corpora information for abbreviation resolution.

Corpus	Medstract	100-Medlines
Abstracts	133	100
Abbreviations	168	162
LF(Ab)/NP(NP)	163/298	162/380
Ab(LF)/NP(NP)	2/298	0/380
NP, NP,	3/143	0/162
(NP)NP	0/0	0/0

In Table 2, characters beside alphabet and number are used to separate words into tokens.

Table 2: Number of tokens in long forms and NP-chunks collect form Medstract.

Abb. length in characters (n)	LF length in tokens (LF)	Extended NPs in tokens (EN)	min(n*2,n+5) number as tokens (MIN)	Pattern Count (P)
2	1	6	4	1
2	1	7	4	1
2	1	5	4	1
2	2	2	4	2
2	2	5	4	5
…	…	…	…	…

$$\frac{\sum_{1}^{N}(|EN-LF|)\times P}{\sum_{1}^{N}P} \quad v.s. \quad \frac{\sum_{1}^{N}(|MIN-LF|)\times P}{\sum_{1}^{N}P} = 3.7 \quad v.s \quad .4.8$$

The average of difference between extended NPs in tokens and long form length in tokens is 3.7. The average of difference between 'min(n*2,n+5)' number as tokens and long form length in tokens is 4.8. The result shows that using extended NPs as candidate of long form is closer to right long form chunk, so we need to delete less words.

We use precision, recall and F-Score to evaluate our result, precision and recall function is listed below:

$$recall = \frac{\#of\ \ pairs\ \ correctly\ \ identified}{\#of\ \ correct\ \ \ pairs} \tag{1}$$

$$precision = \frac{\#of\ \ pairs\ \ correctly\ identified}{\#of\ identified\ \ pairs} \tag{2}$$

We use chunker result as our baseline, the baseline model is using base NP definition, patterns such as ‘NP1 of NP2’ is considered as NP1 and NP2. But from Table 3 we can see the short definition NPs can not cover most long form, so the long form chunker is using the extended NPs, patterns such as ‘NP1 of NP2’ is considered

as one NP.

On the Medstract corpus, our method results in precision 94% at a recall of 86%. For comparison, the algorithm described in Schwartz and Hearst [03] achieved 96% precision at 82% recall, and that of Pustejovsky et al.[01] achieved 98% precision at 72% recall.

Table 3: Experimental results on 100-Medlines w.r.t. rules.

	100-Medlines				Medstract		
Threshold	Type of NP	Recall	Precision	F-Score	Recall	Precision	R-Score
Chunker Result	Base-NP	66.05%	72.30%	69.03%	53.57%	63.38%	58.06%
All	Extended-NP	85.80%	99.29%	92.05%	85.12%	94.08%	89.38%
All-R5	Extended-NP	**88.89%**	**97.96%**	**93.20%**	**86.90%**	**95.42%**	**90.97%**
All-R5-R1-R3	Extended-NP	7.41%	48.00%	12.83%	4.17%	77.78%	7.91%
All-R5-R2	Extended-NP	**88.89%**	**97.96%**	**93.20%**	**86.90%**	**95.42%**	**90.97%**
All-R5-R3	Extended-NP	56.79%	94.85%	71.04%	48.21%	90.00%	62.79%
All –R5-R4	Extended-NP	82.72%	91.16%	86.73%	63.10%	92.98%	75.18%
All-R5-R6	Extended-NP	88.89%	97.30%	92.90%	82.14%	94.52%	87.90%
All-R5-R7	Extended-NP	85.19%	97.87%	91.09%	82.14%	94.52%	87.90%

Table 3 shows the impact factor for each rule. It is noticed that rule 1 and rule 3 carry more information in long form identification.

In 100-Medlines results, we extract ten error abbreviation and long form pairs. Six are because the first character is not the same with the first character of abbreviation, one is because order of abbreviation and long form character is not same, and three are not relation abbreviation and long form pairs. Amount unsolved 18 abbreviation and long form pairs, three are NP chunking error, 13 are syllables error, and one is <Heliothis receptor 14-16, HR14 HR16>. To compare with base line model, we found out chunker can correctly chunk instances which contain semantic type of

abbreviation pair such as < RNA polymerase II, Pol II> pair but for instance such as <differentiation inhibitory factor,I factor> will chunk the long form error, because NP will contain 'differentiation' while long form only contain 'inhibitor factor'.

Table 4: Experiments on 100-Medlines for various threshold.

	Abbreviation and Chunk Correct		
Threshold	Recall	Precision	F-Score
100%	85.80%	99.29%	92.05%
90%	85.80%	99.29%	92.05%
80%	85.80%	99.29%	92.05%
70%	88.27%	98.62%	93.16%
60%	**88.89%**	**97.96%**	**93.20%**
50%	90.12%	90.68%	90.40%

Table 5: Experiments on Medstract corpus in various thresholds.

	Abbreviation and Chunk Correct		
Threshold	Recall	Precision	F-Score
100%	82.14%	94.52%	87.90%
90%	82.14%	94.52%	87.90%
80%	82.74%	94.56%	88.25%
70%	**86.31%**	**94.16%**	**90.06%**
60%	85.71%	91.72%	88.62%
50%	85.12%	89.38%	87.20%

Table 4 and Table 5 show results in different threshold, and we can see the best result is about the same threshold (66%) in the Pustejovsky et al. [01]. It means the match ratio is at 66% between abbreviation and its corresponding long form.

Table 6 shows count of each rule is fired in Medstract, it shows the most important rules are R1, R2, R3, R4 and R7. Most R2 cases can be done by other rules.

Table 6: Rules used in Medstract.

	R1	R2	R3	R4	R5	R6	R7
Medstract	168	67	56	36	4	1	22

Chapter 3.

Anaphora Resolution

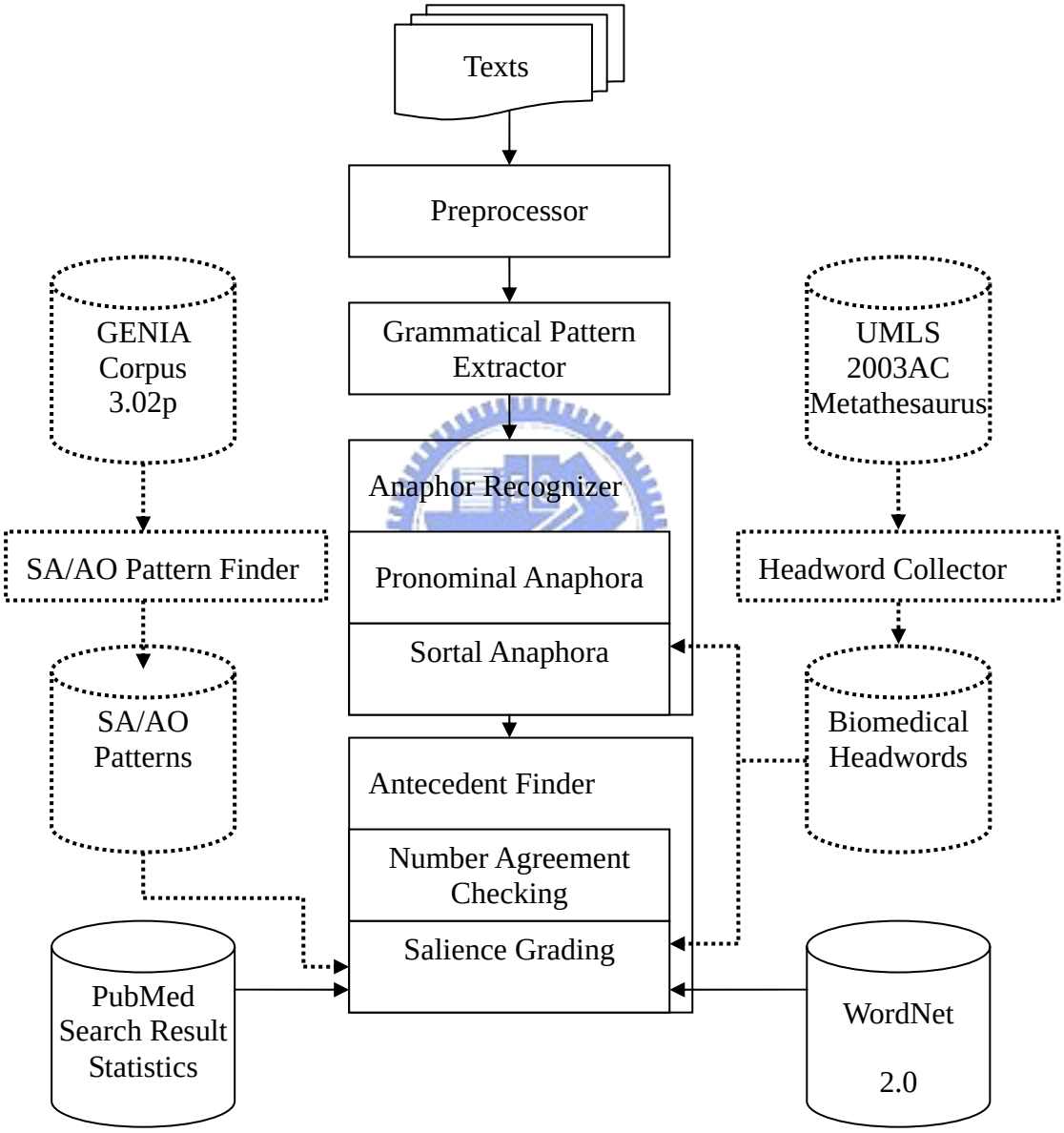


Figure 2: Architecture overview.

Figure 2 is the presented overview architecture which contains background processing indicated with dotted lines, including biomedical SA/AO patterns and headword

collection and foreground processing indicated with solid lines, including preprocessor, grammatical pattern extractor anaphor recognizer, and antecedent finder.

3.1. Headword Collector

For unknown words, we need to predict their semantic types of the word. In [Pustejovsky et al., 02a], they use the right-hand head rule (the head of a morphologically complex word to be the right-hand member of that word) to extract headwords to be subtype of the semantic type in UMLS (135 semantic types).

We collected all UMLS concepts and their corresponding synonyms (1,860,682 recorders), 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. We set the threshold as 0.03, if the maximum words of the semantic is over 10000 the threshold is 0.003.

Headword scoring function:

$$w_{i,j} = \frac{w_i}{Max\ c_j} \times \frac{1}{tw_i} \qquad (4)$$

$w_{i,j}$: 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

Table **7**: Top score headwords for Amino Acid, Peptide, or Protein semantic type.

Headword	Score	No. Count
Protein	0.020833	36807
Product	0.007223	12761
Cerevisiae	0.007082	3128
endonuclease	0.005832	1288
Kinase	0.00575	2963
Antigen	0.004536	4842
Receptor	0.004478	4450
Synthase	0.004426	1629
Reductase	0.004279	1575
Arabidopsis	0.004246	1094
dehydrogenase	0.004005	2064
Antibody	0.003867	3416

3.2. SA/AO Patterns Finder

In this thesis we used co-occurring SA/AO patterns obtained from GENIA corpus for pronominal anaphora resolution. We use the English Part-of-Speech Tagger (http://tamas.nlm.nih.gov/tagger.html) proposed by Tamas Doszkocs, Ph.D to tag POS and NPs, then we use the grammatical function extractor to extract subject and objects. Then we tag subjects and objects with UMLS-semantic type tags, we search the noun phrase from right to find the longest word sequence can found in the UMLS, if not found we will try the headwords to tag semantic types. Each SA/AO pattern is scored by the scoring function (Eq. 1). An antecedent candidate is concerned if its scores are greater than a given threshold (0.01).

$$score(type_i, verb_j) = \frac{frequency(type_i, verb_j)}{frequency(verb_j)} \times \frac{1}{No.\ of\ types(verb_j)} \quad (3)$$

The following is a pattern extraction example:

Example 7:

<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>

Table 8 is a statistic of pattern association with the verb 'bind' and possible semantic type for its subject.

Table 8: Statistics of patterns (subject, verb).

Score (Pharmacologic Substance,	Bind) =	0.142857
Score (Organic Chemical,	Bind) =	0.114286
Score (Amino Acid, Peptide, Protein	Bind) =	0.114286
Score (Cell,	Bind) =	0.085714

3.3. Preprocessor

Anaphor resolution step, we use the tagger which base-NP will be chunked.

3.4. 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], with rules 5 and 6.

Rule 1: Prep NP (Oblique)

Rule 2: Verb NP (Direct object)

Rule 3: Verb $[NP]^{+}$ NP (Indirect object)

Rule 4: NP (Subject) [“,[^Verb], ”|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 8 will not be found when resolving ‘they’’s antecedents.

Example 8:

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

3.5. 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. In this thesis, we deal with pronominal and sortal anaphor. In current version, zero and event anaphora are not solved.

3.6. 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 such anaphors can be resolved by using two simple rules.

Rule 1: The nearest noun phrase of prepositional phrase is assigned as antecedent of the anaphor."

Rule 2: If the anaphor is 'that' and paired with pleonastic-it, the relative clause next to the anaphor is its antecedent.

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 9.

Table 9: 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

3.6.1. Pronominal Anaphora Recognition

Pronominal anaphora recognition is done by filtering out pleonastic-it. We reference Tyne and Wu [04] and generate following rules are used to recognize pleonastic-it instances.

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

Example 9: “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 10: “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 11: “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 12: “Furthermore, the same experimental model makes it possible to image lymphoid progenitors in fetal and adult hematopoietic tissues.”

3.6.2. Sortal Anaphora Recognition

Sortal anaphora recognition is done by filtering those sortal anaphors, 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 non-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.

3.7. 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.

3.8. Salience Grading

Salience grade for each candidate antecedent is assigned according to Table 10. 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 10: Salience grading for candidate antecedents.

Features	Score
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
Subject and Object Preference	1
Grammatical function agreement	1
Number Agreement	1
Longest Common Subsequence	0 to 3
Semantic Type Agreement	-1 to +2
Biomedical antecedent preference	-2 if not or +2

$$Score\ (A_i) = F_1 + F_2 + F_3 + F_4 + F_5 + F_6 + F_7 \quad (3)$$

3.8.1. 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 anaphors 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: we search the noun phrase from right to find the longest word sequence can found in the UMLS.

(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

Queries are used to query from PubMed website and Eq. 4 is used to grade the

antecedent for semantic type agreement.

$$Score(A_i) = Score(A_i) - 1 + \left\lceil \frac{\#of\ \ pages\ \ containing(Ana, A_i)}{\#of\ \ pages\ \ containing(A_i)} \times 10 \right\rceil \times 0.3 \qquad (4)$$

3.8.2. 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 an anaphor and its candidate antecedents

then salience score = salience score + 3

Else if partial match between an 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

Example 13: total match:

<anaphor: each **inhibitor**, Antecedent: PAH alkyne metabolism-based **inhibitor**s>

Example 14: partial match:

<Anaphor: both **receptor** types, Antecedent: the ETB **receptor** antagonist BQ788>

Example 15: using WordNet 2.0:

<Anaphor: this protein (has hyponym: growth **factor)**, Antecedent: Cleavage and polyadenylation specificity **factor** (CPSF)>

3.8.3. 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.

3.8.4. Feature Selection

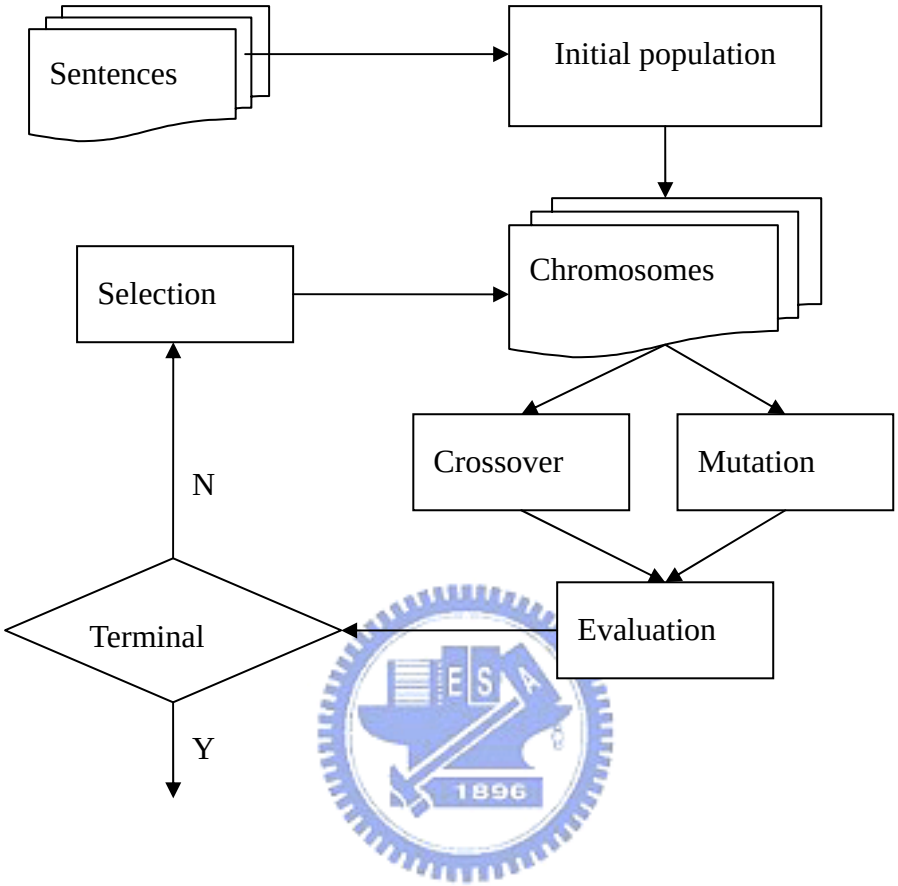


Figure 3: A general of genetic algorithm flowchart.

From Ng and Cardie [2002a] they showed the improvement in F-Score with hand-selected features. Feature selection in this thesis for salience grading is implemented with a genetic algorithm which can get the best features by choosing best parents to produce offspring leave local maximum by mutation. Sequential Floating Forward Selection (SFFS) is the best among the sequential search algorithms, but between the SFFS and GA, no clear cut case can be made for which is the better of the two. [Oh et al., 04].

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. Maximal F-Score is used to evaluate each chromosome and top 10 chromosomes are chosen for next generation. The algorithm terminated if two contiguous generations does not increase the F-score. Time complexity is O(MN) where M is the number of candidate antecedents, N is number of anaphors.

3.9. Experimental Results and Analysis

The test corpus, Medstract, was adopted from (http://www.medstract.org/), containing 32 MEDLINE abstracts and 83 biomedical anaphora pairs (40 pronominal (14 which) and 43 sortal pairs). We try to establish a corpus containing as many kinds of anaphor types as possible, so we collected 43-Genia and 57-Medlines from different ways. We combine 43-Genia and 57-Medlines as 100 MEDLINE abstracts (100-Medlines). 43 abstracts (479 sentences) were from GENIA corpus which contain pronominal anaphor, 57 abstracts (656 sentences) are from PubMed query result by using queries "these proteins" and "these receptors") containing 177 pronominal anaphora and 186 sortal anaphora pairs. Table 12 shows the statistic of pronominal and sortal anaphors for each corpus.

From Table 12 we have number of each anaphor distribution in each corpus. For pleonastic-it we total find 13 instances which all can be resolved. There are 314 'the NP' sortal anaphor candidates in Mestract, 611 'the NP' instances in 43-GENIA, and 607 'the' anaphor candidates in 57-Medlines.

Table 11: Statistics of anaphor and antecedent pairs.

	Abstracts	Sentences	Pronominal instances	Sortal instances	Total
57-Medlines	57	565	69	118	187
43-GENIA	43	479	98	63	161
Medstract	32	268	26	57	83

Table 12: Occurrences of each anaphor.

Pronominal	it	its	itself	they	their	them	themselves	Total
Medstract	6	9	0	2	7	2	0	26
43-Genia	17	42	0	9	13	0	0	84
57-Medlines	10	12	0	7	31	2	0	62
Sortal	the	this	these	those	both	either	each	Total
Medstract	9	9	6	2	11	2	4	43
43-Genia	25	7	11	0	9	0	0	54
57-Medlines	13	5	58	0	12	0	1	89

Table 4 to Figure 5 presents the distribution of the NPs between antecedents and anaphors. From Table 4 and Table 5 we can conclude that NPs between anaphors and antecedents in sortal anaphora are more than NPs in sortal anaphora. Sortal anaphors contain more information than pronominal anaphors, so it is more readable than pronominal anaphors in far distance.

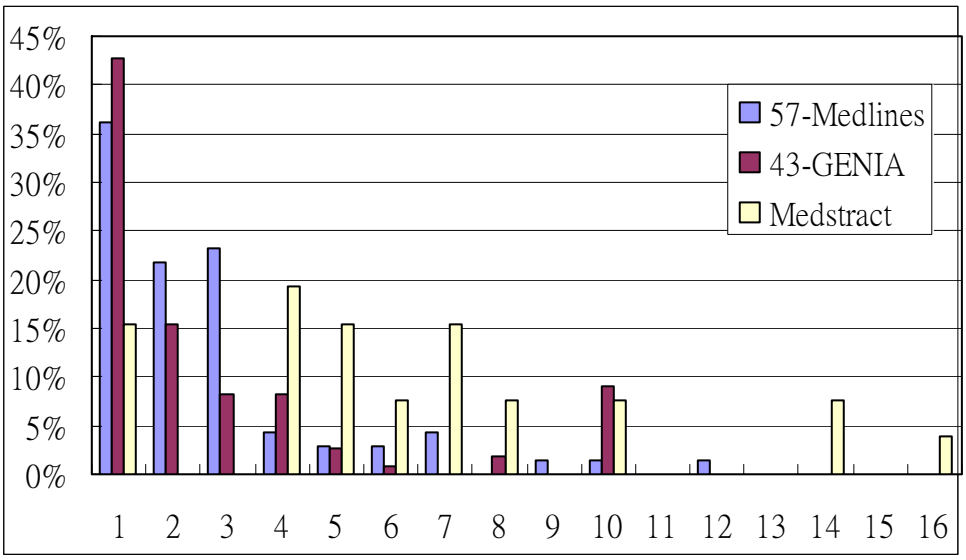


Figure 4: NPs between pronominal anaphor and antecedent.

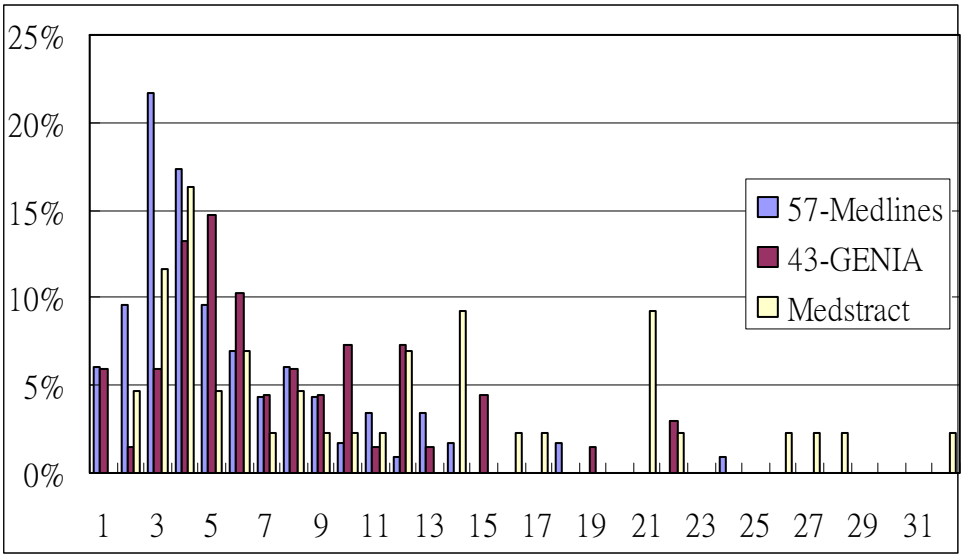


Figure 5: NPs between sortal anaphor and antecedent.

Figure 4 and Figure 5 shows percentage of NPs between anaphor and antecedent in both pronominal and sortal anaphors. From Figure 4 we can see the tendency of fewer instances as the distance increase in pronominal anaphora, wile Figure 5 the highest percentage is not the nearest NPs.

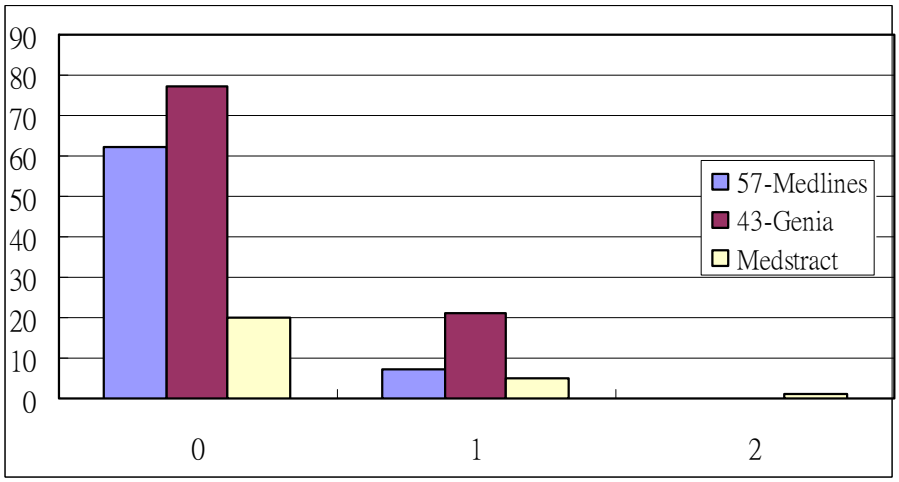


Figure 6: Sentences between pronominal anaphors and antecedents.

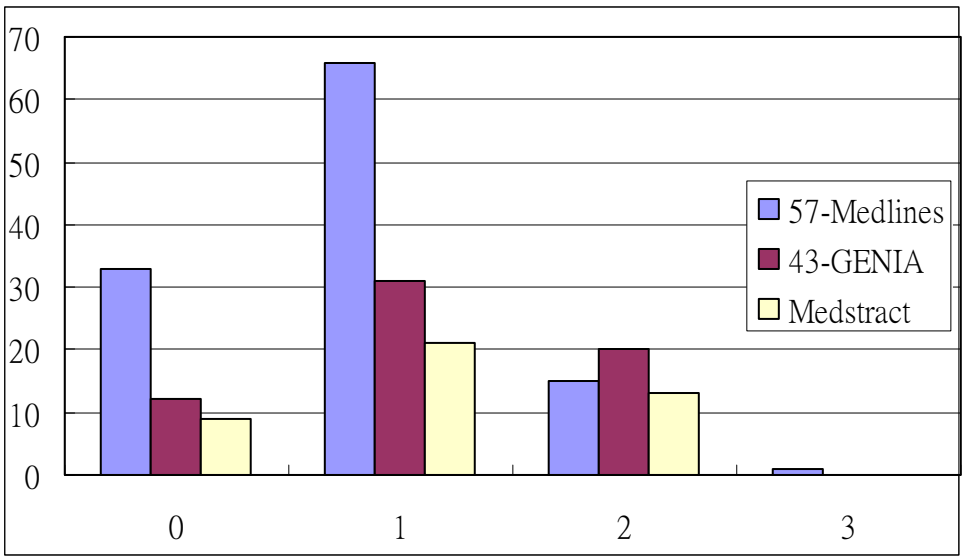


Figure 7: Sentences between sortal anaphors and antecedents.

Figure 6 and Figure 7 shows distance in sentences between anaphor and antecedent. The value 0 denotes intra-sentential anaphora and number 1, 2, 3 indicate inter-sentential anaphora which antecedent is 1, 2 or 3 sentences ahead of anaphor. The results give us confident while using two sentences as searching space.

From the experimental results in Table 13, 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 13: Results with best-first and nearest-first algorithms 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

Table 14: F-Score of Medstract and 100-Medlines

	Medstract		100-Medlines	
	Sortal	Pronominal	Sortal	Pronominal
Total Features	64.08%	88.46%	71.33%	86.65%
Genetic Features	F5~F7	All-{F5}	F5~F7	All-{F5}
	78.26%	92.31%	80.62%	87.25%

The impact of each feature was also concerned and verified within different corpora. Results are showed in Table 15 and Table 16 . 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 16 shows F3 increases F-score in pronominal anaphora but drop F-score in sortal anaphora. Medstract and 100-Medlines results show semantic type match is important in both sortal and pronominal anaphora. Table 17 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. From Table 16 shows that SA/AO collection corpus affects the F-Score within 43-GENIA and 57-Medlines. We collect SA/AO patterns from GENIA corpus, so we can identify semantic type more correctly than in 57-Medlines.

Table 15: Impact of each feature in Medstract and 100-Medlines.

	Medstract		100-Medlines	
	Sortal	Pronominal	Sortal	Pronominal
All	64.08%	88.46%	71.33%	86.65%
All – F1	61.05%	73.08%	70.11%	80.36%
All – F2	65.96%	88.00%	75.55%	86.65%
All – F3	72.00%	80.77%	72.14%	81.30%
All – F4	64.65%	81.48%	71.85%	85.45%
All – F5	48.00%	92.31%	55.18%	87.25%
All – F6	44.04%	88.46%	53.41%	80.24%
All – F7	38.26%	59.26%	55.83%	63.18%

Table 16: Impact of each feature in 43-GENIA and 57-Medlines.

	43-GENIA		57-Medlines	
	Sortal	Pronominal	Sortal	Pronominal
All	67.69%	93.58%	73.28%	76.81%
All - F1	60.14%	83.87%	75.44%	75.36%
All - F2	70.22%	93.58%	78.40%	76.81%
All - F3	69.68%	84.46%	73.45%	76.81%
All - F4	68.33%	91.54%	73.73%	76.81%
All - F5	52.55%	93.58%	56.59%	78.26%
All - F6	46.42%	81.63%	57.14%	78.26%
All - F7	47.19%	71.96%	60.44%	50.72%

Table 17: Impact of headword and PubMed in Medstract.

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

The success rate is calculated as following equation:

$$Success\ Rate = \frac{number\ of\ correctly\ resolved\ anaphors}{number\ of\ all\ anaphors} \quad (6)$$

Success rate shows the accuracy of identifying anaphor and its antecedent. From Table 18, the success rates of sortal anaphora are higher than their F-Score, while success rate of pronominal anaphora are lower than their F-Score. Results shows in 100-MEDLIINES, sortal anaphora have more plural anaphora errors and pronominal have more singular anaphora errors.

Table 18: Success rates of the 100-Medlines.

	100-Medlines	
	Sortal	Pronominal
All	77.30%	82.64%
All – Recency (F1)	77.30%	77.78%
All - Subject or Object preference (F2)	80.85%	78.47%
All - Grammatical Role Match (F3)	76.60%	75.00%
All - Number Agreement (F4)	75.89%	79.86%
All – LCS (F5)	59.57%	82.64%
All – Semantic Type Match (F6)	60.74%	79.17%
All - Biomedical Antecedent (F7)	61.70%	58.33%

Table 19 shows features used in Medstract. From table, the pronominal anaphora use syntax and semantic features except 'F5'. For sortal anaphora, the syntax features are used in selecting antecedent, but from Table 15 and Table 16 we can see that using these feature will drop F-scores in antecedent selection.

Table 19: Features used in Medstract.

	Sortal	Pronominal
F1	30	25
F2	22	15
F3	3	11
F4	21	25
F5	17	0
F6	41	5
F7	37	22

Chapter 4. Conclusion

In this thesis, pronominal and sortal anaphora which are common phenomenon in biomedical texts are discussed. Pronominal anaphora was dealt by using syntactic and semantic features, while sortal anaphora tackled by using semantic features. For new biomedical entities to UMLS, we solved the entities semantic agreement by using headword mining and patterns mined from PubMed query results. Experiment results showed that the proposed strategies indeed enhance the resolution in terms of higher F-Score. For abbreviation resolution, we used more features and gain more F-Score in recognizing the long-form.

Main error to anaphora recognition is that semantic type checking for new words is still lower in precision; in future we may use a larger database such as SWISS-Port for more information for protein or a NE tagger to gain the precision of semantic type check.

References

[1] C. Aone and S. W. Bennett. (1995). “Evaluating Automated and Manual Acquisition of Anaphora Resolution Strategies,” *In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics.*

[2] Breck Baldwin. (1997). “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.

[3] C. Cardie and K. Wagstaff. (1999). "Noun Phrase Coreference as Clustering," *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, 82-89, Association for Computational Linguistics, 1999.*

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

[5] Ido Dagan and Alon Itai. (1999). “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.

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

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

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

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

[10] Jacquemin C. and Tzoukermann E. (1997). "NLP for term variant extraction: Synergy between morphology, lexicon, and syntax." *In: Strzalkowski T., ed, Natural Language Processing and Information Retrieval. Kluwer, Boston, Mass, 1997..*

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

[12] Seth Kulick, Mark Liberman, Martha Palmer, and Andrew Schein. (2003). “Shallow Semantic Annotation of Biomedical Corpora for Information Extraction,” *Proceedings of the 2003 ISMB Special Interest Group Meeting on Text Mining (a.k.a. BioLink).*

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

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

[15] J. McCarthy and W. Lehnert. (1995). “Using DecisionTrees for Coreference Resolution,” *In Proceedings ofthe Fourteenth International Conference on Artificial Intelligence.*

[16] Ruslan Mitkov. (1998). “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.*

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

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

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

[20] Natalia N. Modjeska, Katja Markert and Malvina Nissim. (2003). “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.*

[21] Vincent Ng. (2004). "Learning Noun Phrase Anaphoricity to Improve Coreference Resolution: Issues in Representation and Optimization," *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL), Barcelona, Spain, July 2004, pp. 152-159.*

[22] Vincent Ng and Claire Cardie. (2002a). "Improving Machine Learning Approaches to Coreference Resolution," *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2002.*

[23] Vincent Ng and Claire Cardie. (2002b) "Identifying Anaphoric and Non-Anaphoric Noun Phrases to Improve Coreference Resolution," *Proceedings*

of the 19th International Conference on Computational Linguistics (COLING-2002), 2002.

[24] Vincent Ng and Claire Cardie. (2002). "Combining Sample Selection and Error-Driven Pruning for Machine Learning of Coreference Rules," *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2002*

[25] Il-Seok Oh, Jin-Seon Lee, and Byung-Ro Moon. (2004). "Hybrid Genetic Algorithms for Feature Selection," *IEEE Transactions on pattern analysis and machine Vol. 26, No. 11, 2004*

[26] T. Ohta, Y. Tateisi, J.D. Kim, S.Z. Lee and J. Tsujii. (2001). "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.*

[27] J. Pustejovsky, J. Castaño, B. Cochran, M. Kotecki, M. Morrell, A. Rumshisky. (2001). "Linguistic Knowledge Extraction from Medline: Automatic Construction of an Acronym Database," *An updated version of the paper presented at Medinfo, 2001.*

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

[29] Ariel Schwartz and Marti Hearst. (2003). "Simple Algorithm for Identifying Abbreviation Definitions in Biomedical Text," I*n the proceedings of the Pacific Symposium on Biocomputing (PSB 2003) Kauai, Jan 2003.*

[30] Advaith Siddharthan. (2003). "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). pages 7-14.

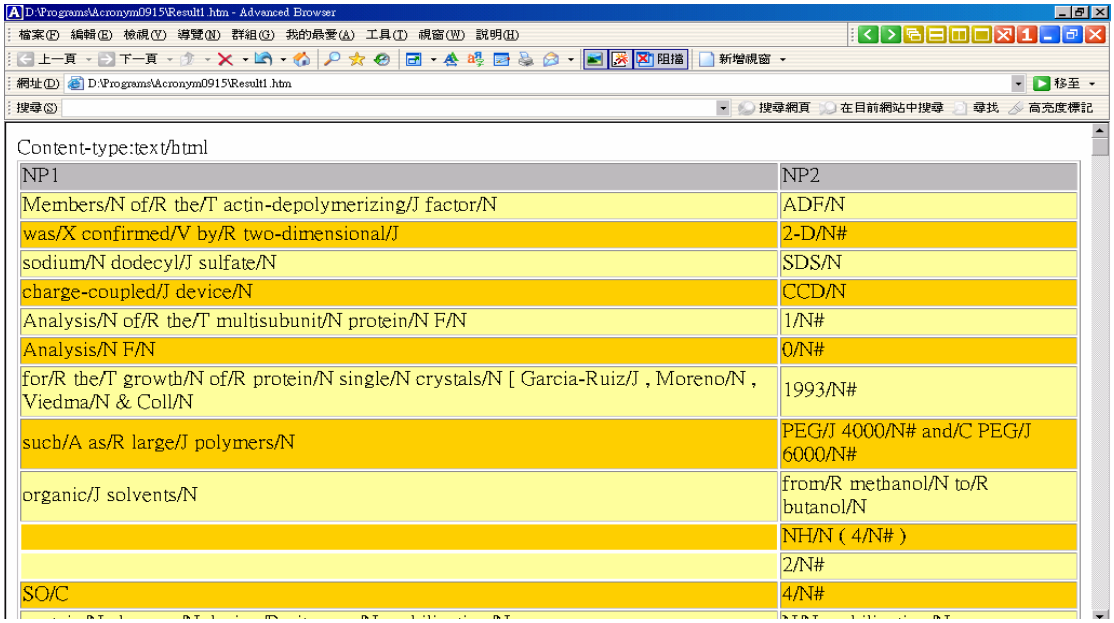
[31] Ping-Ke Shih. (2004). “Automatic Protein Entities Recognition from Pub Med Corpus,” *Thesis. National Chiao Tung University.*

[32] Xiaofeng Yang, Jian Su, Guodong Zhou and Chew Lim Tan. (2004). “Improving Pronoun Resolution by Incorporating Coreferential Information of Candidates,” *In proceedings of ACL 2004, pp.127-134*

[33] M. Yoshida, K. Fukuda, and T. Takagi. (2000). “PNAD-CSS: a workbench for constructing a protein name abbreviation dictionary,” *Bioinformatics, 16(2), 2000.*

Appendix A.

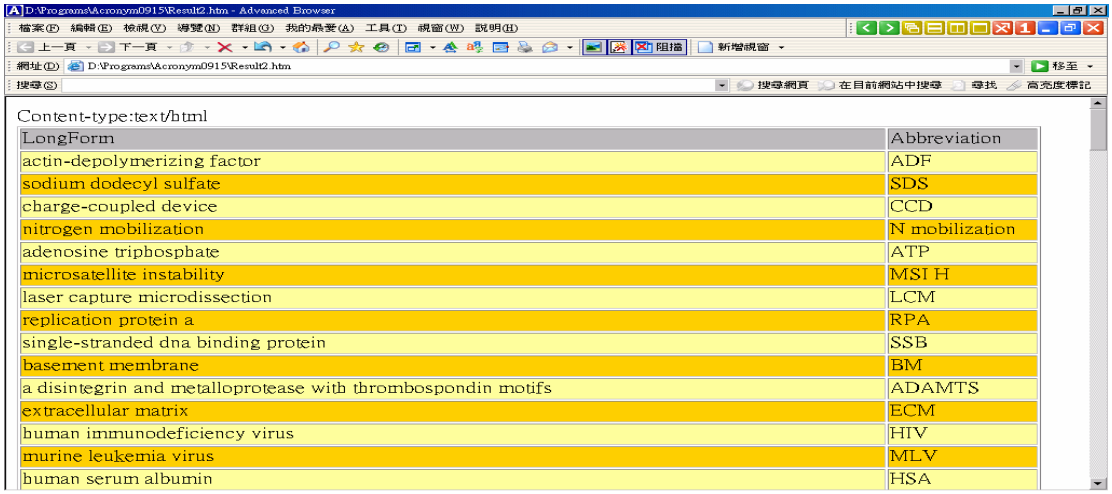
An example of our abbreviation output



Content-type:text/html

NP1	NP2
Members/N of/R the/T actin-depolymerizing/J factor/N	ADF/N
was/X confirmed/V by/R two-dimensional/J	2-D/N#
sodium/N dodecyl/J sulfate/N	SDS/N
charge-coupled/J device/N	CCD/N
Analysis/N of/R the/T multisubunit/N protein/N F/N	1/N#
Analysis/N F/N	0/N#
for/R the/T growth/N of/R protein/N single/N crystals/N [Garcia-Ruiz/J , Moreno/N , Viedma/N & Coll/N	1993/N#
such/A as/R large/J polymers/N	PEG/J 4000/N# and/C PEG/J 6000/N#
organic/J solvents/N	from/R methanol/N to/R butanol/N
	NH/N (4/N#)
	2/N#
SO/C	4/N#

Figure 8: Candidates abbreviation pairs.



Content-type:text/html

LongForm	Abbreviation
actin-depolymerizing factor	ADF
sodium dodecyl sulfate	SDS
charge-coupled device	CCD
nitrogen mobilization	N mobilization
adenosine triphosphate	ATP
microsatellite instability	MSI H
laser capture microdissection	LCM
replication protein a	RPA
single-stranded dna binding protein	SSB
basement membrane	BM
a disintegrin and metalloprotease with thrombospondin motifs	ADAMTS
extracellular matrix	ECM
human immunodeficiency virus	HIV
murine leukemia virus	MLV
human serum albumin	HSA

Figure 9: Abbreviation pairs output.

Appendix B.

An example of UMLS Metathesaurus

UMLS Knowledge Source Server (UMLSKS)

UMLSKS Version 4.2.2 UMLS Releases: 2002 2002AB 2002AC 2002AD 2003AA 2003AB 2003AC 2004AA 2004AB

Metathesaurus Semantic Network SPECIALIST Lexicon

Home Advanced Search Logout

Metathesaurus Search for: **IL-2** in UMLS Release **2004AA**

Concept

☑ Definition

☑ Synonyms

☐ Other Languages

☐ Suppressible Synonyms

☐ Sources

Context

☐ Ancestors

☐ Parents

☐ Siblings

☐ Children

Relations

☐ Narrower

Concept: Interleukin-2

CUI: C0021756

Semantic Type: Amino Acid, Peptide, or Protein

Immunologic Factor

Definition:

A soluble substance elaborated by antigen- or mitogen-stimulated T-LYMPHOCYTES which induces DNA synthesis in naive lymphocytes. (MeSH)

IL-2. A type of biological response modifier (a substance that can improve the body's natural response to infection and disease) that stimulates the growth of certain disease-fighting blood cells in the immune system. These substances are normally produced by the body. Aldesleukin is IL-2 that is made in the laboratory for use in treating cancer and other diseases. (Physician Data Query)

Synonyms:

Interleukin-2

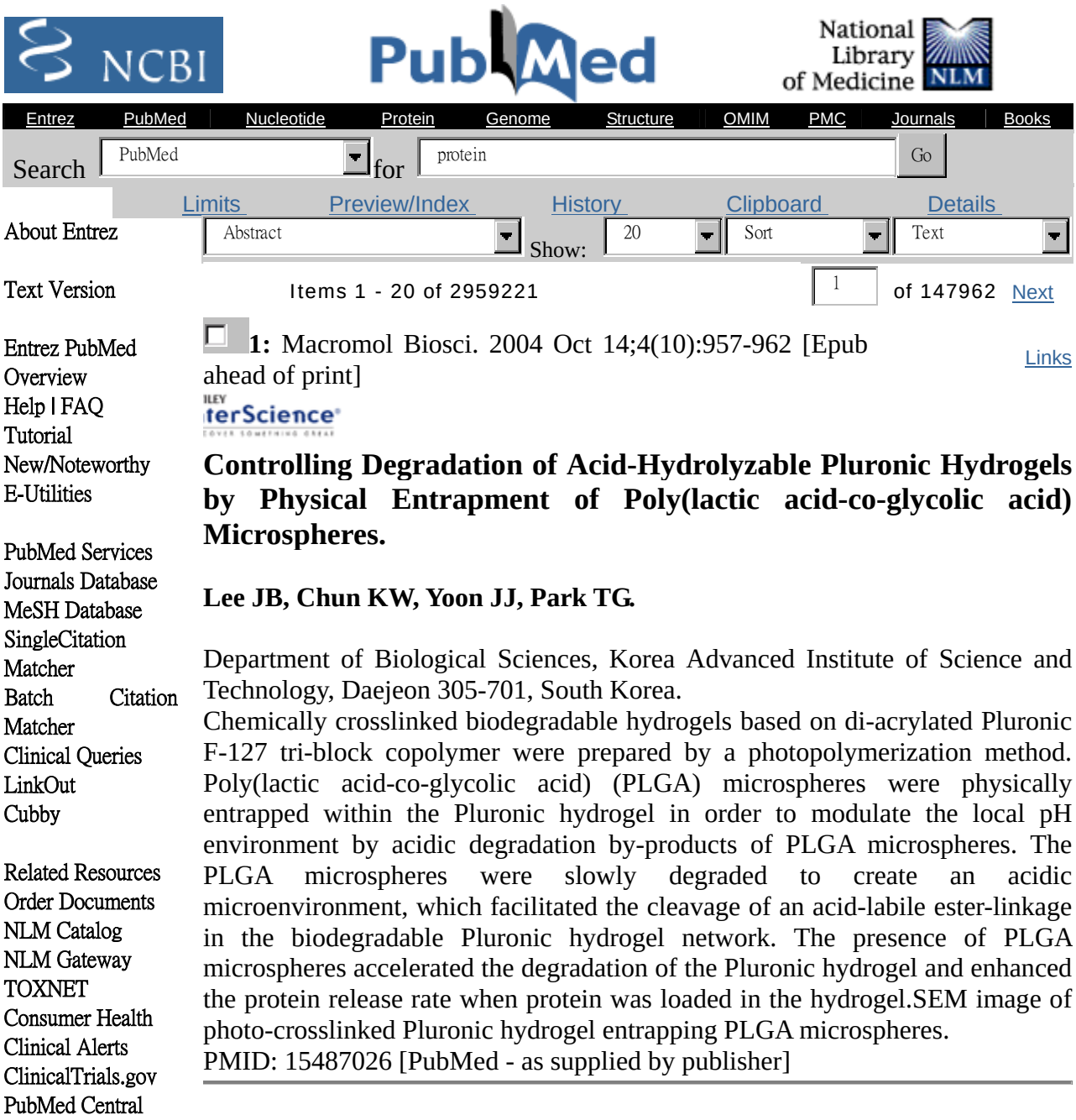
Costimulator

Co-Stimulator

☐ Broader ☐ Similar ☐ Other ☐ Related and possibly synonymous ☐ Source asserted synonymy ☐ Allowable Subheadings ☐ Associated Expressions	IL2 IL-2 Interleukin-2 (substance) Interleukine 2 Interleukin II Lymphocyte blastogenic factor Lymphocyte Mitogenic Factor Lymphocyte mitogenic factor (substance) T-Cell Growth Factor T-Cell Stimulating Factor TCGF TCGF, Interleukin TCGF (T cell growth factor) Thymocyte Stimulating Factor TSF TSF (thymocyte stimulating factor) T-stimulating factor
Co-occurring Concepts	
☐ Co-occurring MeSH	
☐ Co-occurring MeSH By Semantic Group	
☐ Co-occurring AI/RHEUM	
☐ Co-occurring AI/RHEUM By Semantic Group	

Appendix C.

An example of PubMed query result



NCBI PubMed National Library of Medicine NLM

Entrez PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Books

Search PubMed for protein Go

Limits Preview/Index History Clipboard Details

About Entrez

Text Version

Entrez PubMed Overview
Help | FAQ
Tutorial
New/Noteworthy
E-Utilities

PubMed Services
Journals Database
MeSH Database
SingleCitation Matcher
Batch Citation Matcher
Clinical Queries
LinkOut
Cubby

Related Resources
Order Documents
NLM Catalog
NLM Gateway
TOXNET
Consumer Health
Clinical Alerts
ClinicalTrials.gov
PubMed Central

Abstract Show: 20 Sort Text

Items 1 - 20 of 2959221 1 of 147962 Next

1: Macromol Biosci. 2004 Oct 14;4(10):957-962 [Epub ahead of print] Links

Controlling Degradation of Acid-Hydrolyzable Pluronic Hydrogels by Physical Entrapment of Poly(lactic acid-co-glycolic acid) Microspheres.

Lee JB, Chun KW, Yoon JJ, Park TG.

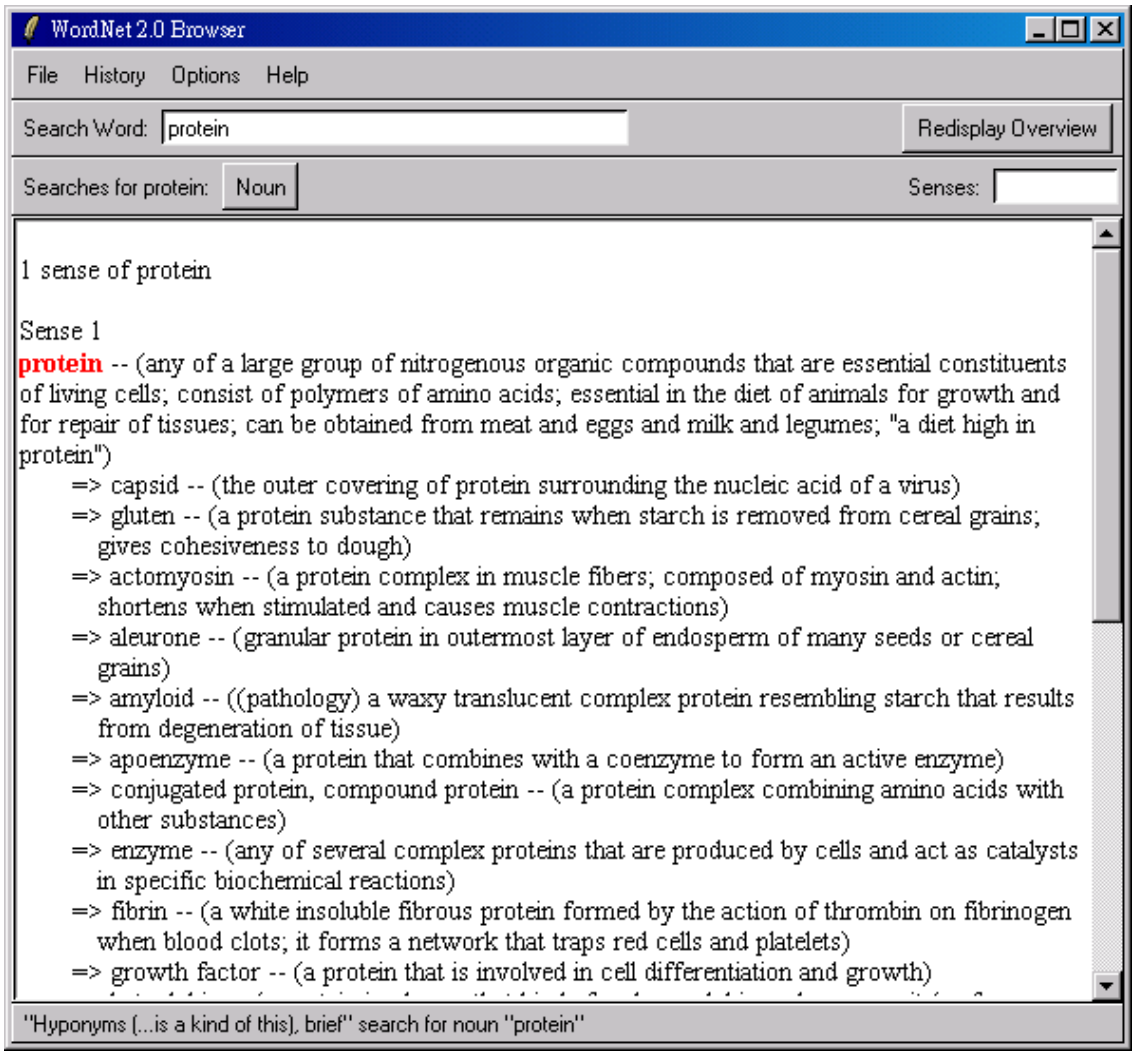
Department of Biological Sciences, Korea Advanced Institute of Science and Technology, Daejeon 305-701, South Korea.

Chemically crosslinked biodegradable hydrogels based on di-acrylated Pluronic F-127 tri-block copolymer were prepared by a photopolymerization method. Poly(lactic acid-co-glycolic acid) (PLGA) microspheres were physically entrapped within the Pluronic hydrogel in order to modulate the local pH environment by acidic degradation by-products of PLGA microspheres. The PLGA microspheres were slowly degraded to create an acidic microenvironment, which facilitated the cleavage of an acid-labile ester-linkage in the biodegradable Pluronic hydrogel network. The presence of PLGA microspheres accelerated the degradation of the Pluronic hydrogel and enhanced the protein release rate when protein was loaded in the hydrogel.SEM image of photo-crosslinked Pluronic hydrogel entrapping PLGA microspheres.

PMID: 15487026 [PubMed - as supplied by publisher]

Appendix D.

An example of WordNet 2.0 result



Appendix E.

An example of GENIA 3.02

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/css" href="gpml.css" ?>
<!DOCTYPE set SYSTEM "gpml.dtd">
<set>

<import resource="GENIAontology.daml" prefix="G"></import>

<article>
<articleinfo>
<bibliomisc>MEDLINE:95369245</bibliomisc>
</articleinfo>
<title>
<sentence><cons lex="IL-2_gene_expression" sem="G#other_name"><cons lex="IL-2_gene" sem="G
</title>
<abstract>
<sentence>Activation of the <cons lex="CD28_surface_receptor" sem="G#protein_family_or_gro
<sentence>In <cons lex="primary_T_lymphocyte" sem="G#cell_type">primary T lymphocytes</con
<sentence>Delineation of the <cons lex="CD28_signaling_cascade" sem="G#other_name"><cons l
<sentence>Our data suggest that <cons lex="lipoxygenase_metabolite" sem="G#protein_family_
<sentence>These findings should be useful for <cons lex="therapeutic_strategies" sem="G#ot
</abstract>
</article>

<article>
<articleinfo>
<bibliomisc>MEDLINE:95333264</bibliomisc>
</articleinfo>
<title>
<sentence>The <cons lex="peri-kappa_B_site" sem="G#DNA_domain_or_region">peri-kappa B site
</title>
<abstract>
<sentence><cons lex="human_immunodeficiency_virus_type_2" sem="G#virus">Human immunodefici
<sentence><cons lex="HIV-1" sem="G#virus">HIV-1</cons> and <cons lex="HIV-2" sem="G#virus"
```

Appendix F.

An example of MEDSTRACT

```
<?xml version="1.0"?>
<!DOCTYPE Articles
    PUBLIC "-//Medstract//DTD XML Medstract annotations//EN//XML"
    "http://www.cs.brandeis.edu/~steele/medstract/Medstract.dtd">
<Articles><Article>
<T>Immunohistochemical localization of <Entity id="5" Type="Protein">FAP-
1</Entity>, an <InhibitRelation id="7" Inhibitor="5"  Inhibitee="6"
Kind="Undetermined">inhibitor of</InhibitRelation> <Entity id="6"
Type="cellular process">Fas-mediated apoptosis</Entity>, in normal and
neoplastic human tissues</T>
<A>S. H. Lee, M. S. Shin, W. S. Park, S. Y. Kim, H. S. Kim, J. H. Lee, S. Y.
Han, H. K. Lee, J. Y. Park, R. R. Oh, J. J. Jang, J. Y. Lee and N. J. Yoo</A>
<J>Apmis</J>
<V>107</V>
<P>1101-8</P>
<Y>1999</Y>
<Ab><Entity id="1" Type="Protein">Fas</Entity>, a death receptor, is widely
expressed in human tissue, but <Entity id="3" Type="molecular process"><Entity
id="2" Antecedent="1">its</Entity> expression</Entity>, although a prerequisite
for the induction of apoptosis, does not predict <Entity id="5"
Type="physiological process"><Entity id="4" Antecedent="2">its</Entity>
biological function</Entity>. To understand the mechanisms of Fas resistance in
human tissues in vivo, we performed immunohistochemistry using an antibody
against <Entity id="8" Type="Protein">Fas-associated phosphatase-1
</Entity>(<Entity id="9" Aliasof="8">FAP-1</Entity>), <Entity id="10"
Antecedent="9">which </Entity>interacts with the cytosolic domain of Fas and
<InhibitRelation id="12" Inhibitor="10"  Inhibitee="11"
Kind="Undetermined">inhibits </InhibitRelation><Entity id="11" Type="cellular
process">Fas-mediated apoptosis</Entity>. In normal human tissues, FAP-1
immunostaining was easily detected, for example, in renal tubules, skeletal
```

Appendix G.

An example of Medstract Gold Standard Evaluation Corpus

```
<?xml version="1.0"?>
<!DOCTYPE Articles
    PUBLIC "-//Medstract//DTD XML Medstract annotations//EN//XML"
    "http://www.cs.brandeis.edu/~steele/medstract/Medstract.dtd">
<Articles><Article>
<T>The leukemic protein core binding factor beta (CBFbeta)-smooth-muscle myosin
heavy chain sequesters CBFalpha2 into cytoskeletal filaments and aggregates</T>
<A>N. Adya, T. Stacy, N. A. Speck and P. P. Liu</A>
<J>Mol Cell Biol</J>
<V>18</V>
<P>7432-43</P>
<Y>1998</Y>
<Ab>The fusion gene CBFB-MYH11 is generated by the chromosome 16 inversion
associated with acute myeloid leukemias. This gene encodes a chimeric protein
involving the <Entity id="1" Type="Protein">core binding factor beta</Entity>
(<Entity id="2" Aliasof="1">CBFbeta</Entity>) and the <Entity id="3"
Type="Protein">smooth-muscle myosin heavy chain</Entity> (<Entity id="4"
Aliasof="3">SMMHC</Entity>). Mouse model studies suggest that this chimeric protein
CBFbeta-SMMHC dominantly suppresses the function of CBF, a heterodimeric
transcription factor composed of DNA binding subunits (CBFalpha1 to 3) and a non-
DNA binding subunit (CBFbeta). This dominant suppression results in the blockage of
hematopoiesis in mice and presumably contributes to leukemogenesis. We used
transient-transfection assays, in combination with immunofluorescence and green
fluorescent protein-tagged proteins, to monitor subcellular localization of CBFbeta
-SMMHC, CBFbeta, and CBFalpha2 (also known as AML1 or PEBP2alphaB). When expressed
individually, CBFalpha2 was located in the nuclei of transfected cells, whereas
CBFbeta was distributed throughout the cell. On the other hand, CBFbeta-SMMHC
formed filament-like structures that colocalized with actin filaments. Upon
cotransfection, CBFalpha2 was able to drive localization of CBFbeta into the
nucleus in a dose-dependent manner. In contrast, CBFalpha2 colocalized with CBFbeta
-SMMHC along the filaments instead of localizing to the nucleus. Deletion of the
```