Toward Information Extraction: Identifying protein names from biological papers

K. FUKUDA, T. TSUNODA, A. TAMURA, T. TAKAGI
Human Genome Center, Institute of Medical Science, University of Tokyo,
4-6-1 Shirokanedai, Minato-ku,
Tokyo 108, JAPAN
e-mail : ichiro@ims.u-tokyo.ac.jp

To solve the mystery of the life phenomenon, we must clarify when genes are expressed and how their products interact with each other. But since the amount of continuously updated knowledge on these interactions is massive and is only available in the form of published articles, an intelligent information extraction (IE) system is needed. To extract these information directly from articles, the system must firstly identify the material names. However, medical and biological documents often include proper nouns newly made by the authors, and conventional methods based on domain specific dictionaries cannot detect such unknown words or coinages. In this study, we propose a new method of extracting material names, PROPER, using surface clue on character strings. It extracts material names in the sentence with 94.70% precision and 98.84% recall, regardless of whether it is already known or newly defined.

1 Introduction

Genome projects are now determining whole DNA sequences and genes of various species. The next target is to analyze the function of each gene. To understand the role of each gene, we must clarify when it is expressed and how its product interacts with other materials. Typical interaction includes protein-nucleotide and protein-protein bindings and many researchers are submitting articles on such kind of interactions. The quality and amount of continuously updated knowledge has become powerful enough to grapple with the mystery of life phenomenon. However, reading every article in the world requires too much time and labor. Therefore, we need an intelligent information extracting system to save time.

To extract information on interactions directly from each article, the system must firstly identify material names, e.g. gene names and protein names. Identifying technical terms from quantities of unrestricted text is a challenging task for natural language processing, especially in medical and biological papers. One will encounter the following difficulties: unknown word processing, long compound word recognition, and requirement of robustness against ambiguous expressions that are used only among the area experts. As far as we know, there is no technical term extracting system that overcome these problems at once.

Typical method to specify technical terms and proper nouns in texts is to match each word to the heading words on prepared dictionaries. However, medical and biological texts typically contain proper nouns newly made by the authors. Hence, one cannot avoid encountering unknown words, and adding such words to the dictionary for future convenience is extremely time consuming and may result in many mistakes. Furthermore, when the word expression in the document is different from the dictionary headings, this conventional method is powerless.

Compound words are common in technical terms. Some compound extraction approaches using statistical methods has been proposed [6,7]. However, the work of K. Su *et al*[6] considers only bigrams and trigrams[a], while medical and biological papers commonly include compounds of more than six words (for example, "Ras guanine nucleotide exchange factor Sos"). Furthermore, ambiguous expressions which are used only among the area experts exist. Hence, it seems difficult to get a discriminative statistical threshold.

Many Information Extraction Systems have been proposed in the Message Understanding Conference(MUC) [3], and the known best strategy is to prepare proper noun dictionaries and a pattern dictionary [b]. The drawbacks are; first, the performance largely depends on the quality of their proper name dictionaries, second, usually one does not have an annotated corpus to learn patterns, third, preprocessing is necessary so that patterns can extract compound words as one word.

To cope with the above three problems, we propose a new method of extracting material names such as proteins, using surface clue on character strings in medical and biological documents. This method, PROPER (PROtein Proper-noun phrase Extracting Rules), uses the characteristics of proper noun description in these fields, and does not require any specific term dictionary prepared in advance. It extracts material names in the sentence with high accuracy regardless of whether it is already known or newly defined, and whether it is a single word or a compound word. As a result, our method can completely correspond to the variation in expression.

2 Features of protein names

2.1 Nomenclatures of protein names

Protein names can be classified into the following three categories from their structure.

[a] Their recall and precision are 96.2% , 48.2% for bigrams and 96.6%, 39.6% for trigrams

[b] An example of patterns in terrorism article is "killed ⟨ dobj ⟩" [8], where the name of the victim to be extracted comes in ⟨ dobj ⟩

1. Single words with upper case letters, numerical figures, and non-alphabetical letters. Mostly derived from gene name (*ex.*Nef, p53, Vav)
2. Compound words with upper case letters, numerical letters, and non-alphabetical letters. (*ex.*interleukin 1 (IL-1)-responsive kinase)
3. Single word with only lower case letters (*ex.*actin, tublin, insulin)

In the field of medicine and biology, new findings of materials and functions are frequently reported. When new material is discovered, the researcher usually creates a new term which can be clearly distinguished from other materials and concepts. Therefore, in actual papers, protein names of type 3 is relatively rare, whereas material names like type 1 and 2 are often observed. These material names are usually new words which are introduced by the researcher and are updated daily.

2.2 Variation in Expression

In addition to newly introduced words, variation and inconsistency in referring already known materials is another serious problem. Especially when mentioning a material whose name explains its role, the expression is almost arbitrary. Also, single word proper nouns may have inconsistency in its spelling.

1. Authors often use the original words instead of abbreviations, change letter cases, and ignore implicit name generating rules.

   - epidermal growth factor receptor *or* EGF receptor *or* EGFR
   - cycline D1-cdk4 complex *or* cycline D1-Cdk4 complex
   - c-Jun *or* c-jun *or* c jun

2. Below, the name explains its function.

   - the Ras guanine nucleotide exchange factor Sos
   - the Ras guanine nucleotide releasing protein Sos
   - the Ras exchanger Sos
   - the GDP-GTP exchange factor Sos
   - Sos(mSos), a GDP/GTP exchange protein for Ras

   They show that the description of protein names in the field of medicine and biology is extremely variant.

3. Next examples include preposition and/or conjunction. Due to the ambiguity of dependencies, the variation in description can be more complex compared to that in the second examples.

    - p85 alpha subunit of PI 3-kinase
    - SH2 and SH3 domains of Src

Thus, description of material names often depends on the author's style, and there is no guarantee that the same protein will appear in the same description in different sentences and documents.

2.3 Features of protein names

In spite of this arbitrariness of protein name description, a significant characteristic exists in technical terms of this field. Characteristic words containing capital letters, numerical figures, and special symbols as underlined in the following examples are frequently observed. These words can be clearly distinguished from general words.

- Src homology (SH) 2 and SH3 domains
- p54 SAP kinase

These words provide large amount of information to the reader and can be considered as the core of material names. In this respect, we call such words that appear in protein names "core-terms."

Furthermore, as in the following example, key-words can be included that describe the function and characters of compound word.

- EGF receptor
- Ras GTPase-activating protein (GAP)

We will call such words "f-terms(feature-terms)."

By focusing on these characteristics, it will become easier to find the candidates of material names, including those that are newly introduced.

3 methods

In this study, we have extracted the following as "target material names."

- Protein name (including kinase, receptor, ligand, enzyme, compound)
- Protein domain name or motif, site, fragment, and element, etc. which are narrower regions represented by specific sequences

It is extremely important to specify these material names in order to automatically extract the findings of cell signal transduction from documents, and to establish databases by an Information Extraction (IE) system.

The target material name is extracted by the following two phases.

1. Core-term extraction from tokenized texts.
2. Concatenation of core-terms and f-terms.
   (a) rebuild "core-blocks" [noun-phrases without conj and prep]
   (b) rebuild dependencies

This will be explained in detail hereafter.

3.1 Extraction method of core-term

We have extracted core-terms by five serial processings. The first processing extracts all words that are syntactically predicted to be a "core-term", and passes the obtained result to the next processing. Filters 2 to 5 remove those that are semantically unacceptable as core-terms from the candidate words obtained in filter 1. The following are the specific contents of the processings.

1. Extract words with upper cases, numerical figures, and/or special symbols as candidates for core-term

2. Exclude words whose length is more than 9 characters and consists of "-" and lower cases. By this process, words such as "full-length" or "dual-specificity" that are used other than in target material names can be removed

3. Exclude words in which more than half of its character string consists of special symbols. This process eliminates character strings such as "+/-".

4. Exclude words related to numbers such as units. Eight words (aa, AA, fold, bp, nM, microM, %, UV) which were registered beforehand as "units" and those ending with these units are removed.

5. Exclude words that agree to the reference template prepared beforehand. As a result, the name of persons and journals in references is removed from the core-term candidates. The reference template is implemented by regular expression.
   example: (Z. Weng, J. A. Taylor, C. E. Turner, J. S. Brugge, and C. Seidel-Dugan, J. Biol. Chem. 268 :14956-14963 , 1993)

3.2 Concatenation of core-terms

Extracted core-terms are annotated in the text sentences. The annotations are extended to adjacent words or concatenated with other annotations. By this process, noun-phrases without conjunctions and prepositions (we call them "core-blocks") are restored. Then, dependencies between these core-blocks are rebuilt.

rebuild core-blocks

We have used the following concatenation rules for core-terms and f-terms. The underlined words in the left hand side of each arrow have already been annotated.

1. only by surface clue

   (a) Annotation is simply connected when core-term or f-term is adjacent to each other.
   Src SH3 domain → Src SH3 domain

   (b) Parentheses are annotated in the following cases.

      i. (SH3) → (SH3)
      ii. (SH2 (and|or) SH3) → (SH2 (and|or) SH3)

2. using POS tagger
   In the next three rules, part of speech information obtained by Brill POS tagger[1] was used as an application condition.

   (a) Connect non-adjacent annotations if every word between them are either noun, adjective, or a numeral.
   Ras guanine nucleotide exchange factor Sos
   → Ras guanine nucleotide exchange factor Sos

   (b) Extend the annotation to the left if there is a determiner or preposition.
   the focal adhesion kinase (FAK) → the focal adhesion kinase (FAK)

   (c) Extend the annotation to the right if there is a single upper case letter or a word representing greek letter.
   p85 alpha → p85 alpha

rebuild dependencies

Since we need to determine dependencies only within noun-phrases, the number of possible combinations is lower compared to that within a full-sentence. Therefore, the ambiguity is reduced, and we achieved in establishing the rebuilding rule with several simple patterns. The rule we used will be explained with examples hereafter. In the following example, A, B, C, D, and E represent core-blocks.

1. "A, B, ... C and D f-term"
   Src , Fyn , Lyn , Yes , and PI3K SH3 domains
   → Src , Fyn , Lyn , Yes , and PI3K SH3 domains
2. "A, B, ... C and D of E"
   Src homology 2 (SH2) and 3 (SH3) domains of Vav
   → Src homology 2 (SH2) and 3 (SH3) domains of Vav
3. "A of B, C and E"
   SH2 domains of Abl , Lck , Fyn , and p85
   → SH3 domains of Abl , Lck , Fyn , and p85
4. "A f-term core-term and core-term"
   GTP-binding proteins Rac1 and Cdc42
   → GTP-binding proteins Rac1 and Cdc42
5. "A of B"
   p85 alpha subunit of PI 3-kinase
   → p85 alpha subunit of PI 3-kinase
6. "A , B"
   the Src-related tyrosine kinase , Hck
   → the Src-related tyrosine kinase , Hck

demark unnecessary annotation

Next, improper annotation must be excluded. Though satisfactorily high recall can be obtained by the above-mentioned processings, they do not possess a rule that corrects wrong annotations. We added two rules. First rule fires when the annotated f-term is not extended and remains as a single word. This is caused since f-terms are very ordinary words. Second rule fires when the last word of the phrase obtained by extension-concatenation procedure is not a noun. This is caused since core-term is not always a noun, as in the case of "Src-related." Annotation is removed or shifted in both rules by pattern matching of regular expressions.

We obtained excellent results for both recall and precision by the above-mentioned rule. The results obtained by this method will be discussed in the next section.

4 experiments

We evaluated the above-mentioned processing by 30 abstracts on SH3 domain (SH3) and 50 abstracts on signal transduction (SGN). All abstracts were retrieved from MEDLINE.

Table 1 shows the amount of f-terms and extracted core-terms in target material names. This result shows that about 95% of target material names

text	target name	inc. c-terms	ratio	inc. f-term or c-term	ratio
SH3	689	623	90.42%	661	95.93%
SGN	749	653	87.20%	705	94.12%

Table 1: The rate of core-term and f-term in target material names; c-term:core-term

always include either a core-term or a f-term. The remaining 5% are words like "insulin," "adenylyl cyclase," and "dynamin."

4.1 Evaluation of core-term extraction phase

Table 2 is the result of core-term extraction phase. All false-positives [c] arise

text	core-term	extracted	f-p	p-f	precision	recall
SH3	198	208	3	15	92.74%	98.48%
SGN	231	230	6	11	95.22%	97.40%

Table 2: result of core-term extraction phase ; f-p:false-positive,p-f:positive-false

in the second processing. "interleukin-beta " is an example of such false-positives. This occurs because the rule could not syntactically distinguish them from words like "full-length". Cell names and virus names were observed in positive-falses [d].

4.2 evaluation of concatenation phase

We classified the errors as follows. 1 and 2 are positive-false, whereas 3 is false-positive.

1. Errors in annotation site(unsu:unsuitable)

[c] terms that could not be extracted
[d] terms that were extracted as an error

(a) Not a protein name
"NINS" (abbreviation of "the 258-bp novel insert")

(b) Represent material name but were excluded from target material name in this study
"PC12 cell", "filamentous bacteriophage fuse5"

(c) Not a specific material name
"major tyrosine-phosphorylated protein"

2. Errors in concatenation & extension processing

(a) Incomplete extension
" interleukin 1 (IL-1) -responsive kinase"

(b) Over-extension
" *same* proline-rich region of FAK (APPKPSR)"

(c) Incomplete connection of preposition and conjunction
" p80 and p85 (p80/85)"

3. Failure in annotation (f-p:false-positive)
"insulin", "adenylyl cyclase"

According to the above classification, the evaluation of the concatenation phase resulted as shown in Table 3. Here, the total appearances of errors were counted. Therefore, if an error in certain annotation appeared three times in a single text, it was counted as three errors. In evaluation 1, all target material

text/measure	ANNO	AUTO	f-p	unsu	disag	p-f	precision	recall
SH3/1	689	683	40	26	24	59	91.90%	93.32%
SH3/2	646	679	1	26	24	59	91.31%	99.85%
SH3/3	646	663	1	10	24	43	93.51%	99.85%
SGN/2	669	666	19	17	43	65	90.24%	97.16%

Table 3: result ; p-f:unsu + disag

names are included as the object of evaluation. In evaluations 2 and 3, the target material names consisting of only lower cases were excluded from the evaluation. Evaluation 3 is when extraction was allowed for cell names and phage names.Target material names with wrong POS tags were removed from the object of evaluation in SGN.

The errors classified as type 1 in the previous section (cell names and phage names) can be distinguished from target material names by recognizing

the words surrounding it("in PC12 cell"). Therefore, extracting these terms will not become a noise in the Information Extraction phase.

Our evaluations have strictly counted the errors in restoration of dependencies(error type 2.(c)). For instance, although all core-blocks are correctly annotated in the following example, since restoration of dependencies was incomplete, five positive-falses were counted.

experimental result:

Grb2, Crk, Abl,p85 phosphatidylinositol 3-kinase, *and* GTPase-activating protein SH2 domains

In natural language processing, restoration of dependencies is generally a very difficult problem due to its ambiguity. Although the uncertainty can be reduced by limiting the range to noun phrases, it was not possible to solve the problem completely.

In order to eliminate errors classified as type 3, new rules for core-term extraction is needed. In Table 4, the following rule is included as a core-term extraction rule.

X is a core-term if it matches ".*'consonant'(in | ase | ol)s{0,1}"

This rule is an expansion of the core-term definition, and words such as "insulin" and "phospholipase" will be newly added as "core-terms".

Evaluation 4 in Table 4 is similar to evaluation 1 in that all target material names are considered as objects of evaluation, except that cell names and phage names were considered as acceptable.

text/measure	ANNO	AUTO	f-p	unsu	disag	p-f	precision	recall
SH3/4	689	698	8	11	21	37	94.70%	98.84%

Table 4: best result ; p-f:unsu + disag

From the comparison of SH3/1 and SH3/4, it is apparent that f-p has been reduced. Also, unsu+disag decreased by 32, while total decrease in unsu and disag was 18. This is due to successful restoration of dependencies.

5 Discussion

Our technique applies to all expressions except some words that lack surface clues. That is, it applies to new words, coinages, long compound words, and variations in expression, which are weak points of conventional methods. In this section, we discuss two problems in extracting terms that have poor surface clue. One is to be solved as a future work and the other has been solved by adding rules as described before.

5.1 How to reduce errors classified as type 1(c)

"focal adhesion kinase" and "major tyrosine-phosphorylated protein" both contain a f-term, and neither of them contain a core-term. Hence, if one extracts the former by the extension-connection rule described previously, he/she cannot avoid extracting non-specific material names like the latter one. Thus, precision must be sacrificed to some extent in order to improve recall.

However, compound material names are often abbreviated, as in "FAK" for "focal adhesion kinase". In this case, the abbreviation is usually defined at the beginning of the text, such as "focal adhesion kinase (FAK)." Since our method can extract such abbreviation-defining paraphrases, the trade-off between precision and recall can be solved by extracting synonyms from such expressions and feeding them back to the text.

5.2 How to reduce errors classified as type 3

There are two options. In words like "adenylyl cyclase" and "insulin", surface clues exit, such as the usage of consonants and vowels and their ending patterns. We actually extracted these words by using additional rules, with the obtained result of 94.70% precision and 98.84% recall.

It should be noted that in these phrases, new words are rarely observed as mentioned before and that they lack variation in expression. Therefore they can be extracted easily by a dictionary prepared in advance. This suggests that if an excellent dictionary is available and if its usage improves the accuracy of our technique, utilizing it together with our technique can be an alternative.

6 Future work

Feeding back the extracted knowledges, phrases, or sub-phrases to the text can be useful in gaining higher precision. Combining prepared dictionaries skillfully to the output of our technique can be helpful to gain better recall.

In our experimental results, scores for SGN was relatively lower than that of SH3. It is assumed that over-fitting of the applied rules to SH3-related documents is the cause. In order to improve the generality, it is necessary to perform experiments with increased number of documents. Moreover, the level of recall and precision achieved by applying our technique in full texts should be confirmed.

Extracting protein names from texts is one problem and identifying protein names that refer to the same material is another. The latter problem remains to be solved. Nomenclature of protein names is "whimsy" and names that are totally different in syntax can refer to the same material [9]. This hampers the comprehension of the continuously updated knowledge. Hence a flexible

automatic-dictionary construction system will be needed as well as a protein-interaction extracting system.

7 Conclusions

We have proposed for the first time a method called PROPER that extracts target material names by performing connection and extension processing around the core-term, using surface clues alone. By this method, we have succeeded in obtaining material names without utilizing any background knowledge.

Our method achieves a recall of 98.84% and a precision of 94.70%. It can satisfactorily bear practical use as a preprocessing for the information extraction task from the document.

Acknowledgments

This work was supported in part by a Grant-in-Aid for Scientific Research on Priority Areas, "Genome Science," from the Ministry of Education, Science, Sports and Culture in Japan.

References

1. E. Brill, "Some Advances in Transformation-Based Part of Speech Tagging" in *Proc. of the Twelfth National Conference on Artificial Intelligence(AAAI-94).*
2. R. Grishman, "The NYU System for MUC-6 or Where's the Syntax" in *Proc. of sixth Message Understanding Conference.(MUC-6).*
3. MUC-6 *Proc. of sixth Message Understanding Conference.(MUC-6).*
4. T.Wakao, R. Gaizauskas, Y Wilks, "Evaluation of an Algorithm for the Recognition and Classification of Proper Names" in *Proc. of the 16th International Conference on Computational Linguistics(COLING 96).*
5. D. A. Evans, C. Zhai, "Noun-Phrase Analysis in Unrestricted Text for Information Retrieval" in *Proc. of the 34th Annual Meeting of the Association for Computational Linguistics(ACL'96).*
6. K.Su, M. Wu, J. Chang, "A Corpus-based Approach to Automatic Compound Extraction" in *Proc. of the 32th Annual Meeting of the Association for Computational Linguistics(ACL'94).*
7. F. Smadja, "Retrieving Collocations from Text: Xtract" in *Computational Linguistics Volume 19, Number 1, 1993.*
8. E. Riloff, "Automatically Generating Extraction Patterns from Untagged Text" in *Proc. of the 13th National Conference on Artificial Intelligence(AAAI-96).*
9. "Obstacles of nomenclature", *Nature Volume 389 Issue 6646 1997.*