Automatic acquisition and classification of terminology using a tagged corpus in the molecular biology domain*

Nigel Collier, Chikashi Nobata and Junichi Tsujii

This article describes our work to identify and classify terms in the domain of molecular biology according to examples that have been marked up by a domain expert in a corpus of abstracts taken from a controlled search of the Medline database. Automatic acquisition of biomedical term lists has so far been slow due to high variability in both the terms and their classification scheme, which we attribute to the diversity of research disciplines involved. Nevertheless, the explosive growth in online molecular biology literature makes a persuasive case for automating many tasks. This includes acquisition of records for gene-product databases such as SwissProt which are currently updated by human experts, a task that is both time consuming and often highly idiosyncratic. In this article we report results from a tool based on a hidden-Markov model for extracting and classifying terms that can be used as a key component in an information extraction system. We discuss the results in light of lexical, syntactic and semantic properties of terms that were revealed by our study.

Keywords: information extraction, named entity, molecular biology

1. Introduction

Within the molecular biology community there is a widespread demand for tools that can provide intelligent access to the latest results that are reported in the vast and rapidly growing collections of scientific literature. Such tools, based on information extraction (IE) technology could for example show protein reactions and help researchers understand signaling events taking place within and between cells in the human body. An IE tool for this task would

Terminology 7:2 (2001), **239–257.**
ISSN 0929–9971/ E-ISSN 1569–9994

depend heavily on the ability to identify and classify the names of genes and gene products that appear in the electronic forms of journal papers and abstracts.

In this article we discuss and present the results of our initial work into developing a tool for term identification and classification from raw text using automatic learning methods based on hidden Markov models (HMMs), a widely used paradigm that has been used in natural language processing (NLP) for syntactic word classification tasks, and in IE for the named entity task in the Message Understanding Conferences (MUCs) (MUC7 1998). This work was conducted as part of the GENIA project (Collier et al. 1999) at the University of Tokyo. In particular, we report here on our analysis of the nature of the task itself through insights gained from our test set, a corpus of 100 abstracts that were retrieved from Medline (PubMed)[1] in a controlled way and then annotated by domain experts according to a top-level ontology (conceptual hierarchy) that our group developed based on chemical substances.

We consider that it is necessary to take a multi-domain view of IE in order to simplify what is already a complex task. For example consider the difficulty that a general domain IE system would have with classifying *tax activation,* where *tax* is a protein in molecular biology and a payment to a government in economics. A multi-domain view for us means developing generic tools that can easily be adapted to new domains without necessarily sharing knowledge between domains.

If an effective tool can be developed for term extraction, it can be used as the basis for many higher-level IE tasks that could potentially be used to automate the updating of gene databases such as SwissProt (Bairoch et al. 1997) or Genbank at NCBI (National Center for Biotechnology Information) and so help researchers to have access to the latest results in a structured form. Terminology lists could also play an important role in helping to integrate information between related databases.

2. Background

Within the NLP community as a whole there has recently been growing interest in the opportunities offered to IE by molecular biology and a number of approaches have been considered. Here we briefly review some of the major projects and then comment on difficulties that are particular to term extraction in molecular biology.

Thomas et al. (1999) adapted Highlight, a general purpose IE system that uses quite shallow linguistic analysis based on cascading finite state techniques to the acquisition of protein interactions in Medline. By analyzing approximately 200 Medline abstracts they were able to customize their system to the domain. In tests on 2565 unseen abstracts Thomas et al. were able to achieve precision in the range of 69 to 77 per cent. and recall in the range of 29 to 55 per cent. Although their work focuses on a higher level IE task than the one we look at here, it incorporates term extraction technology although it is not clear from their analysis just how this affects the overall result.

Sekimizu et al. (1998) in a preliminary investigation considered verb subcategorization information from shallow parsing as a precursor to term extraction and categorization.

Craven et al. (1999) seek to accomplish a number of tasks with their system, including updating databases, summarization and helping in scientific discovery. Initially they concentrate on identifying passages in text that contain a number of relations that are of interest to molecular biologists such as sub cellular-localization and tissue-localization. In their experiments they compare two approaches for sentence identification. The first is an adaptation of the Naive Bayes approach, and the second uses shallow parsing to learn syntactic relation rules. Both models were tested on 2889 Medline abstracts and the relation-rule model was found quite low recall (21 per cent).

Rindflesch et al. (1999) present results of an IE pilot study into automatically creating a high coverage database of molecular binding relations. In experiments they create a test set of 116 Medline abstracts from a search using terms related to binding of proteins. 346 sentences were found that contained a form of *bind* and the authors tried to identify the verb arguments by drawing on existing resources at the National Library of Medicine (NLM) such as the UMLS (Unified Medical Language System) Metathesaurus, SPECIALIST lexicon and GenBank. For unknown words they used simple orthographic heuristics to guess the class of words. The results showed that approximately 72 per cent. of 1064 binding terms could be recalled at 79 per cent. precision.

Considering the term extraction task itself, in our work we consider that from the perspective of expandability, portability as well as transparency, the latest tools based on machine learning paradigms such as HMMs (Bikel et al. 1997), decision trees (Sekine et al. 1998) and maximum entropy (Borthwick et al. 1998) offer significant advantages over those based on hard-wired heuristic rules such as (Fukuda et al. 1998) (Wacholder et al. 1997), shallow linguistic analysis, e.g. Tersmette et al. (1988), and traditional gazette-based methods used

in early studies. Advances in smoothing techniques (Chen et al. 1996) have made it possible to overcome previous problems associated with data sparseness. By improving accuracy at this lowest level of IE task, we hope to improve the performance of higher level IE systems that depend on the results of this technology.

We should not however underestimate the difficulty of this task, particularly for molecular biology. Given the huge number of new gene products now being discovered by ongoing sequencing projects it is unsurprising that no established naming system is being used. This is despite the best efforts of major journals to standardize and gain consensus. We have found that in some cases researchers use common names for gene products that are often formed from whatever name they find meets their intuition. For example, the recently reported *yotiao* protein (Westphal et al. 1999) that occurs in the brain, named by a team after a Chinese breakfast noodle. Clearly if such naming methods become common then we have to expect that IE must work much harder to identify and classify terms. Using the analysis of McDonald (1996) we may consider that evidence for identifying and classifying a term can be based on term-internal evidence (provided by the words in the term itself) or external context (possibly including world knowledge). The successful analysis of *yotaio* requires dependence on external evidence. Named entity techniques, such as the one we present in this article, tend to incorporate a rather weak form of external evidence from within the same sentence (in our method it is imposed both by n-grams and a high probability path through the lattice of term-class probabilities for the sentence). Although we will not discuss this further here, there are cases where external evidence within a sentence may only be adequate for identification but not for classification and *coreference* between named entities needs to be established. An example of this is in the case of an abbreviation where the full form of the term mentioned at the beginning of a text provides all the necessary internal evidence to classify it. Less obviously are cases we found in the molecular biology domain where a term can appear as more than one class in the same text such as a DNA and a protein that was derived from that DNA. In such cases deeper knowledge about the domain is necessary.

In the design stage of our system we considered the need for a flexible tool that would conform to the intuitions that the domain experts have about terms rather than impose on them boundaries and classes for terms that they may not necessarily agree with. Therefore it is important that the user's knowledge and intuitions should be represented naturally and formally. In our case we asked a domain expert who has a doctorate in molecular biology to create a corpus of examples which was then used as training data by our system.

To help in this task we encouraged the domain expert to provide a simple ontology describing the hierarchical structure between term classes. For molecular biology it was decided (Tateishi et al. 1999) to base the structure of the ontology on chemical substance relations rather than the roles of terms which can so often be ambiguous leading to multi-class terms.

Clearly there is a balance to be struck between high accuracy and speed of processing and system development. Although we will see that the HMM we present later performs with reasonable accuracy and high speed, it nevertheless inevitably suffers where there is structural ambiguity. We found very early in our study that the disadvantage of our approach is that the complex interaction between term semantics and syntax meant that we had difficulty separating the functions for term boundary identification and classification. This was exacerbated by limitations in the term markup scheme which imposed non-nesting of terms — reflected in a disjoint term concept hierarchy. We discuss this further in the next section.

3. Domain analysis

Our study was conducted on a small subset of Medline abstracts that were found using a controlled search of the database using the keywords *human, blood cell, transcription factor*. This search produced over 3300 abstracts from which we selected 100 for semantic tagging using a human expert in the domain of molecular biology (Ohta et al. 1999). Although the expert had no formal training in computational linguistics or terminology, she had considerable knowledge about term boundaries and classes gained from reading and research experience so that we considered her annotations to be our 'gold standard', even though they sometimes called for unsatisfactory choices to be made. For easy interchange of data with other groups we chose to use a simple bracketed notation as the basis for markup.

On average each abstract contains about 11 lines including an identification code, a title and the body of the abstract itself. We decided not to use author names, institution name or any other lines that are available in Medline abstracts in this study as they do not contain terms of interest to us can easily be detected and omitted. An example of a single marked-up abstract can be seen in Figure 1. Term open and close boundaries are indicated with '[' and ']' respectively, followed by the term's semantic class given in subscript. Where no subscript follows, the '[' and ']' characters were part of the original text.

Using the analysis of Thomas et al. (1999) we consider a name to belong to any of the following types: proper names, compound nouns containing proper names, compound nouns not containing proper names, a definite description linked to a proper name, and any other name. We did not however consider a name to be a definite description not linked to a proper name.

UI — 92371447
TI — [TAR]$_{RNA}$-independent transactivation by [Tat]$_{protein}$ in cells derived from the [CNS]$_{source.ti}$: a novel mechanism of [HIV-1]$_{source.vi}$ gene regulation.
AB — The [Tat protein]$_{protein}$ of [human immunodeficiency virus type 1]$_{source.vi}$ ([HIV-1]$_{source.vi}$) is essential for productive infection and is a potential target for antiviral therapy. [Tat]$_{protein}$, a potent activator of [HIV-1]$_{source.vi}$ gene expression, serves to greatly increase the rate of transcription directed by the viral promoter. This induction, which seems to be an important component in the progression of acquired immune deficiency syndrome (AIDS), may be due to increased transcriptional initiation, increased transcriptional elongation, or a combination of these processes. Much attention has been focused on the interaction of [Tat]$_{protein}$ with a specific RNA target termed [TAR]$_{RNA}$ ([transactivation responsive]$_{RNA}$) which is present in the leader sequence of all [HIV-1]$_{source.vi}$ mRNAs. This interaction is believed to be an important component of the mechanism of transactivation. In this report we demonstrate that in certain [CNS-derived cells]$_{source.ct}$ [Tat]$_{protein}$ is capable of activating [HIV-1]$_{source.vi}$ through a [TAR]$_{RNA}$-independent pathway. A [Tat-responsive element]$_{DNA}$ is found upstream within the viral promoter that in [glial- derived cell lines]$_{source.cl}$ allows transactivation in the absence of [TAR]$_{RNA}$. Deletion mapping and hybrid promoter constructs demonstrate that the newly identified [Tat-responsive element]$_{DNA}$ corresponds to a sequence within the viral [long terminal repeat]$_{DNA}$ ([LTR]$_{DNA}$) previously identified as the [HIV-1 enhancer]$_{DNA}$, or [NF-kappa B domain]$_{DNA}$. DNA band-shift analysis reveals [NF-kappa B]$_{protein}$ binding activity in [glial cells]$_{source.ct}$ that differs from that present in [T lymphoid cells]$_{source.ct}$. Further, we observe that [TAR]$_{RNA}$-deleted mutants of [HIV-1]$_{source.vi}$ demonstrate normal [late gene]$_{DNA}$ expression in [glial cells]$_{source.ct}$ as evidenced by syncytia formation and production of [viral p24 antigen]$_{protein}$.(ABSTRACT TRUNCATED AT 250 WORDS)

Figure 1. Example Medline abstract marked up with named-entities. The original text is from Taylor, J., R. Pomerantz, O. Basgasra, M. Chowdhury, J. Rappaport, K. Khalili, & S. Amini. 1992. "TAR-independent transactivation by Tat in cells derived from the CSN: a novel mechanism of HIV-1 gene regulation", *EMBO J.* 1992 Sep., 11(9), 3395–3403.

During analysis of the corpus we found that a number of syntactic phenomena caused potential complications in the acquisition of terms. Semantic and syntactic ambiguity of NEs have been outlined for the news domain by for example (Wacholder et al. 1997) but we consider it useful to review it here for the molecular biology domain. Broadly speaking the major problems we found

Table 1. Named entity classes. # indicates the number of tagged terms in the corpus of 100 abstracts

Class	#	Example	Description
PROTEIN	2125	JAK kinase	proteins, protein groups, families, complexes and substructures
DNA	358	IL-2 promoter	DNAs, DNA groups, regions and genes
RNA	30	TAR	RNAs, RNA groups, regions and genes
SOURCE.cl	93	leukemic T cell line Kit225	cell line
SOURCE.ct	417	human T lymphocytes	cell type
SOURCE.mo	21	Schizosaccharomyces pombe	mono-organism
SOURCE.mu	64	mice	multi-organism
SOURCE.vi	90	HIV-1	viruses
SOURCE.sl	77	membrane	sub-location
SOURCE.ti	37	central nervous system	tissue
UNK	–	tyrosine phosphorylation	background words

can be divided into coordination, apposition and abbreviation, although there are many other issues that we cannot cover here such as use of negatives in term names such as 'non-T-cells' and the need to infer some term's classes from a domain model. The three major issues are now discussed below.

3.1 Coordination

Coordination, as applied to the appearance of terms in molecular biology texts appears quite frequently through the use of the coordinators 'and', slash('/'), and hyphen ('-'). Hyphenation is particularly troublesome for our implementation of the HMM as it is also one of the orthographic features used in many protein and gene names.

Firstly, (1) gives the most common case where coordination involves no term abbreviation. A"…" is used to indicate that what follows (or preceeds) is omitted as irrelevant to the point being discussed.

(1) Activation of $[\text{JAK kinases}]_{\text{protein}}$ and $[\text{STAT proteins}]_{\text{protein}}$ by $[\text{interleukin-2}]_{\text{protein}}$ and $[\text{interferon alpha}]_{\text{protein}}$,…

In contrast, (2), (3) and (4) both show the need for structural analysis to take place to transform (at least internally within the IE software) the term back to its base form. Comparing (2) and (3) we see that '/' is sometimes used as part of

the term and sometimes as a coordinator.

(2) …like the [c-rel]$_{protein}$ and [v-rel]$_{protein}$ (proto)oncogenes.

(3) …regulated by members of the [rel/NF-kappa B family]$_{protein}$

(4) …involves phosphorylation of several members of the [NF-kappa B]$_{protein}$/[I kappa B protein]$_{protein}$ families.

Finally, (5) and (6) show more complex cases where the annotation scheme has not allowed the domain expert to fully express her intuition about the classes of terms. In (5) we see that the expert was not happy to markup ' [c-]$_{protein}$' as a term without being able to show the attachment to '-rel' and was also uneasy about marking up 'c- and v-rel' as the term. In (6) we see that although the head noun *regions*, which dominates the list should impose a protein category on 'TATA', without a way of marking up this relation, the expert prefers to tag 'TATA' as DNA according to the class of the basic term under discussion. Cases such as these indicate the need for richer markup methods.

(5) This protein reduces or abolishes in vitro the DNA binding activity of wild-type proteins of the same family ([KBF1]$_{protein}$/[p50]$_{protein}$, c- and [v-rel]$_{protein}$

(6) … indicated that multiple regulatory regions including the enhancer, [SP1]$_{protein}$, [TATA]$_{DNA}$ and [TAR]$_{protein}$ regions were important for [HIV]$_{source.vi}$ gene expression.

3.2 Apposition

To quote from Greenbaum et al. (1993): appositions are "Two or more noun phrases are in apposition when they have identity of reference". In the examples we give below, the referent provides useful information about the apposition phrases' class that requires this relation to be recognised. For example in (8), *transcription factor* provides the information that "NF-Kappa B" is a type of protein, and in (10), *retinoblastoma control factor RCF-A* is a *protein complex*. Similarly, the apposition phrases in examples (7) and (9) provide attribute information about the term. This is particularly useful in higher level IE tasks that require us to combine attributes for a particular term in order to build up a mini-CV of the terms' functionality. The challenge posed by appositions is often to know where to start and end the term boundaries, particularly in examples such as (11) where no punctuation is used.

(7) However, similarly to the other [Rel]$_{protein}$-[NF-kappa B]$_{protein}$ complexes, [RelB]$_{protein}$-[p52]$_{protein}$ can upregulate the synthesis of [I kappa B alpha]$_{protein}$.

(8) AB — The transcription factor [NF-Kappa B]$_{protein}$ is stored in the [cytoplasm]$_{source.sl}$…

(9) We have shown earlier that [NF-M]$_{protein}$, the [chicken]$_{source.mu}$ homolog of [C/EBP beta]$_{protein}$, is specifically expressed in…

(10) We have identified a protein complex, the [retinoblastoma control factor A]$_{protein}$ [[RCF-A]$_{protein}$] which binds to the [c-fos retinoblastoma control element]$_{DNA}$ [[RCE]$_{DNA}$] and …

(11) … is stored in the [cytoplasm]$_{source.sl}$ in complexes with the inhibitor protein [I kappa B alpha]$_{protein}$.

3.3 Abbreviation

Abbreviation is a natural process for making the text more concise and readable by preventing repetition of the long forms of names and is used extensively in abstracts to save space. In molecular-biology in particular we have found that texts are replete with abbreviations, making automatic semantic annotation of terms challenging as we cannot use the lexical or other features of their constituent words. The exception to this is the first mention of the term where the abbreviation is defined, e.g. (12), (13) and (14). This will be of particular importance to the future development of our term acquisition method. A discussion of abbreviation formation is unfortunately beyond the scope of this article, but we have built tools for automatically identifying and marking up abbreviations on-the-fly, to be used in later implementations.

(12) The distral enhancer region of the [human immunodeficiency virus type 1]$_{source.vi}$ ([HIV-1]$_{source.vi}$)…

(13) corresponds to a sequence within the viral [long terminal repeat]$_{DNA}$ ([LTR]$_{DNA}$)…

(14) AB — [5-Aminolevulinate synthase]$_{protein}$ ([ALAS]$_{protein}$) catalyzes the first step of the heme biosynthetic pathway.

Of more immediate concern to us are cases where the abbreviation occurs inside the term itself as shown in (15). Here we require deeper analysis than can be obtained through the local contextual view used by the HMM that we will now present.

(15) The [interleukin-2 (IL-2) promoter]$_{DNA}$ consists of several independent [T cell receptor (TcR) responsive elements]$_{DNA}$.

4. Method

Our initial approach has been motivated by the need to acquire the majority of terms that do not involve complex structural analysis to recover their base forms. For this we consider that a HMM approach based on raw text strings of words and no deep linguistic analysis is well suited. Later we envisage that more sophisticated markup and pre-processing methods will be needed to handle term structure and we hope to incorporate these within the automatic learning approach we have adopted in our work.

HMMs themselves can be considered to be stochastic finite state machines and have enjoyed success in a number of fields including speech recognition and part-of-speech tagging (Kupiec 1992). It has been natural therefore that these models have been adapted for use in other word-class prediction tasks such as the named-entity task in IE. Such models are often based on N-gram. Although the assumption that a word's part-of-speech or name class can be predicted by the previous n–1 words and their classes is counter-intuitive to our understanding of linguistic structures and long distance dependencies, this simple method does seem to be highly effective in practice.

Although it is still early days for the use of HMMs for IE, we can see a number of trends in the research. Systems can be divided into those which use one state per class such as Nymble (Bikel et al. 1997) (at the top level of their backoff model) and those which automatically learn about the model's structure such as Seymore et al. (1999).

In the following model, we consider words to be ordered pairs consisting of a surface word, W, and a word feature, F, given as <W, F>. The word features themselves are discussed below. As is common practice, we need to calculate the probabilities for a word sequence for the first word's name class (C_{first}) and every other word differently since we have no initial name-class to make a transition from. Accordingly we use the following equation to calculate the initial name class probability,

Equation 1

$$\Pr(C_{first}|<W_{first},F_{first}>)= \sigma_0 f(C_{first}|<W_{first},F_{first}>)+ \sigma_1 f(C_{first}|<\_,F_{first}>)+ \sigma_2 f(C_{first})$$

and for all other words and their name classes (C_t) as follows:

Equation 2

$$\Pr(C_t|<W_t,F_t>,|<W_{t-1},F_{t-1}>,C_{t-1})= \lambda_0 f(C_t|<W_t,F_t>,|<W_{t-1},F_{t-1}>,C_{t-1})+ \lambda_1 f(C_t|<\_,F_t>,|<W_{t-1},F_{t-1}>,C_{t-1})+ \lambda_2 f(C_t|<W_t,F_t>,|<\_,F_{t-1}>,C_{t-1})+ \lambda_3 f(C_t|<\_,F_t>,|<\_,F_{t-1}>,C_{t-1})+ \lambda_4 f(C_t|C_{t-1})+ \lambda_5 f(C_t)$$

where f(|) is calculated with maximum-likelihood estimates from counts on training data.

In our current system we set the constants λ_i and σ_i by hand and let $\Sigma\sigma_i = 1.0$, $\Sigma\lambda_i = 1.0$, $\sigma_0 \geq \sigma_1 \geq \sigma_2$, $\lambda_0 \geq \lambda_1 \ldots \geq \lambda_5$. The current name-class C_t is conditioned on the current word and feature, the previous name-class, C_{t-1}, and previous word and feature.

Equations 1 and 2 implement a *linear-interpolating* HMM that incorporates a number of sub-models designed to reduce the effects of data sparseness.

Once the state transition probabilities have been calculated according to Equations 1 and 2, the Viterbi algorithm (Viterbi 1967) is used to search the state space of possible name class assignments in linear time to find the highest probability path, i.e. to maximize Pr(W,C).

Currently we have optimized the λ constants by hand but clearly a better way would be to do this automatically. An obvious strategy to use would be to use some iterative learning method such as Expectation Maximization (Dempster et al. 1977). The final stage of our algorithm that is used after name-class tagging is complete is to use a clean-up module called *Unity*. This creates a frequency list of words and name-classes and then re-tags the text using the most frequently used name class assigned by the HMM. We have generally found that this improves F-score performance by between 2 and 4 per cent. both for re-tagging spuriously tagged words and for finding untagged words in unknown contexts that had been currently tagged elsewhere in the text.

4.1 Tokenization

Before featurising we perform two pre-processing tasks: sentence boundary identification and word tokenization. These proceed according to quite simple algorithms that have nevertheless proven to be adequately effective. Sentence boundary identification simply treats full stops'.' as end of sentence except for a few special cases such as abbreviation marking and decimal points that are handled with heuristic rules. Although this method is far less sophisticated than others such as Reynar et al. (1997), we found that it performed well in practice, although we would expect to encounter more significant problems in engineering and technical domains.

Tokenization treats a continuous string of letters and/or numerals as a word, converts multiple space sequences to single spaces, removes non-printable characters except end-of-line, and treats punctuation (including hyphen) as a separate 'word'. Processing is done sentence by sentence. All words are assigned a feature code depending on their orthographic form.

4.2 Orthographic features

In order to generalize the HMMs' knowledge about surface forms it is necessary to featurize the vocabulary in some way. On analyzing lists of terms we felt that orthographic features offered particularly strong clues about the classes of words in molecular biology.

Table 2 shows the character features that we used in the HMM. Our intuition is that such features will help the model to find similarities between known words that were found in the training set and unknown words and so overcome the unknown word problem. Each word is deterministically assigned a single feature, giving matching features nearer to the top of the table priority over those lower down.

5. Experiment and Results

We ran 5-fold cross validation tests on the 100 abstract corpus, taking 80 abstracts for training and 20 for testing. We then calculated the score as an average of the F-scores for each category.

The results are given as F-scores, a common measurement for accuracy in the MUC conferences that combines recall and precision. These are calculated using a standard MUC tool (Chinchor 1995). F-score is defined as:

Equation 3

$$\text{F-score} = (2 \times \text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall})$$

Table 2. Character features with examples. It should be noted that the examples do not show the full form of terms, but simply examples of 'words' that make up the term together with their semantic classification

Feature code	Example(s)
DigitNumber	$[2]_{\text{protein}}$ $[3]_{\text{DNA}}$
SingleCap	$[I]_{\text{protein}}$ $[B]_{\text{protein}}$ $[T]_{\text{source.ct}}$
GreekLetter	$[\text{alpha}]_{\text{protein}}$
CapsAndDigits	$[2A]_{\text{DNA}}$ $[BW5147]_{\text{source.cl}}$ $[CD4+]_{\text{source.ct}}$
TwoCaps	$[RelB]_{\text{protein}}$ $[TAR]_{\text{RNA}}$ $[HMG]_{\text{DNA}}$ $[NF]_{\text{protein}}$
LettersAndDigits	$[p50]_{\text{protein}}$ $[Kit225]_{\text{source.cl}}$
InitCap	$[\text{Interleukin}]_{\text{protein}}$ $[Sox]_{\text{DNA}}$
LowCaps	$[\text{kappaB}]_{\text{protein}}$ $[mRNA]_{\text{RNA}}$
Lowercase	$[\text{cytoplasmic}]_{\text{source.sl}}$ $[tax]_{\text{protein}}$
Determiner	the
Conjunction	and
FullStop	.
Comma	,
Hyphen	$[-]_{\text{protein}}$ $[-]_{\text{DNA}}$
Colon	:
SemiColon	;
OpenParen	(
CloseParen	)
OpenSquare	[
CloseSquare	]
Percent	%
Other	*+#
Backslash	$[/]_{\text{protein}}$

The first set of experiments we did show the effectiveness of the model for all name classes and is summarized in Table 3. We see that data sparseness does have an effect, with proteins — the most numerous class in training — getting the best result and RNA — the smallest training class — getting the worst result. The table also shows the effectiveness of the character feature set, which in general adds 10.6 per cent. to the F-score. This is mainly due to a positive effect on words in the PROTEIN and DNA classes, but we also see that members of all SOURCE sub-classes suffer from featurization.

Returning to the earlier examples, we see that failure occurred where complex structure could not be resolved by the shallow level processing of the HMM. For example in (15) we saw how abbreviations could occur inside terms. The result from the HMM is given in (16).

(16) The $[\text{interleukin-2}]_{\text{protein}}$ ($[\text{IL-2}]_{\text{protein}}$) promoter consists of several independent $[\text{T cell receptor}]_{\text{protein}}$ ($[\text{TcR}]_{\text{protein}}$) responsive elements.

Table 3. Named entity acquisition results using 5-fold cross validation on 100 tagged Medline abstracts, 80 for training and 20 for testing. *Base-features* uses no character feature information

Class	Base	Base-features
PROTEIN	0.759	0.670 (−11.7%)
DNA	0.472	0.376 (−20.3%)
RNA	0.025	0.000 (−100.0%)
SOURCE (all)	0.685	0.697 (+1.8%)
SOURCE.cl	0.478	0.503 (+5.2%)
SOURCE.ct	0.708	0.752 (+6.2%)
SOURCE.mo	0.200	0.311 (+55.5%)
SOURCE.mu	0.396	0.402 (+1.5%)
SOURCE.vi	0.676	0.713 (+5.5%)
SOURCE.sl	0.540	0.549 (+1.7%)
SOURCE.ti	0.206	0.216 (+4.9%)
All classes	0.728	0.651 (−10.6%)

It is important to remember that the HMM looks for the most likely sequence of classes that correspond to the word sequence and that, for example, the word sequence for "interleukin-2" is far more likely to be a protein than a DNA, given that abbreviations usually occur after the term has finished and are mostly marked up as separate terms in their own right. Here though we have an exception and it requires quite sophisticated processing to recognize the embedded abbreviation does not form part of the term itself, and that the head of the first term is "promoter". A similar case can be found in the second term 'T cell receptor responsive elements', where the abbreviation should also be considered as a separate term. The difficulty can be traced back to limitations in our markup scheme which did not allow the domain expert to express her intuition about the term's structure with either nested or cross-over.

Examples such as (3) and (4) also resulted in difficulties as they provide

conflicting patterns, i.e. sometimes '/' should act as a coordinator and sometimes as part of the term itself as can be seen in the HMM output in (17) and (18).

(17) … regulated by members of the rel/[NF-kappa B family]$_{protein}$

(18) … involves phosphorylation of several members of the [NF-kappa B/I kappa B] $_{protein}$ families

Apposition by itself in (7) did not seem to be the cause of the problem, but rather the conjunction implied by the hyphen makes it unclear where to break the sequence "RelB-p52" as seen in (19). Interestingly the HMM managed to correctly find the break in the earlier sequence "Rel-NF-kappa B'.

(19) However, similarly to the other [Rel] $_{protein}$-[NF-kappa B] $_{protein}$ complexes, [RelB-p52] $_{protein}$ can upregulate the synthesis of [I kappa B alpha] $_{protein}$.

The apposition of (8) and (11) also posed no difficulty for the HMM as shown in (20) and (21) respectively.

(20) The transcription factor [NF-Kappa B] $_{protein}$ is stored in the [cytoplasm]$_{source.sl}$…

(21) … is stored in the [cytoplasm] $_{source.sl}$ in complexes with the inhibitor protein [I kappa B alpha]$_{protein}$.

6. Conclusion

Despite a limited context window used in analysis, the HMM performed quite well, showing that finite state techniques can give good results despite shallow linguistic analysis. Unlike traditional dictionary-based term identification methods used in IE, the method we have shown has the advantage of being portable and no hand-made patterns were used. The study also indicated that more training data is better and that we have not yet reached a peak in the level of performance using the small training set that we have available.

Further studies reported elsewhere (Nobata et al. 2000) confirm the importance of lexical and orthographic feature information through analysis of entropies and gain-ratios of features used by this system in two different domains, molecular biology and newswire.

As we would expect, many of the problems were caused firstly by a failure to correctly identify the term boundaries, rather than a failure in classification. This resulted from both genuine failure to find the terms base form and also

confusion in the training data due to inconsistencies in term markup. Markup also forced the expert to ignore intuitions that she may have had about the term's internal class structure. Future work now needs to concentrate on:

– Richer markup methods.[2]
– Examination of the need to model a terms' internal structure at the markup stage.
– Extending the models' contextual view, perhaps by adding the verb and head noun as features.
– Shallow syntactic analysis to reveal term boundaries and complex structures.

For our work to be of most use to molecular biologists corpus markup needs to be extended to cover richer classes of terms such as chemical, drug and disease names.

Notes

* We are grateful to a number of our colleagues at the Tsujii laboratory who contributed comments and support to our research including Yuka Tateishi and Tomoko Ohta for their work on constructing the ontologies and tagged corpora that are used for training in our work and to Sang-Zoo Lee for useful comments on the HMM.

1. Medline is a large database of abstracts and bibliographic information for the bio-medical literature and can be found at http://www.ncbi.nlm.nih.gov/PubMed

2. See Lee et al. (2000) for a technical report discussing extensions to our work in this area.

References

Bairoch, A. and R. Apweiler. 1997. "The SWISS-PROT protein sequence data bank and its new supplement TrEMBL". *Nucleic Acids Research* 25, 31–36.

Bikel, D., S. Miller, R. Schwartz and R. Weichedel. 1997. "Nymble: a high-performance learning name-finder". In *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP'97)*, 194–201, Washington, USA.

Borthwick, A., J. Sterling, E. Agichtein and R. Grishman. 1998. "Exploiting diverse knowledge sources via maximum entropy in named entity recognition". In *Proceedings of the Workshop on Very Large Corpora (WVLC'98)*, 152–160, Montreal, Canada.

Chen, S. and J. Goodman. 1996. "An empirical study of smoothing techniques for language modeling". In *Proceedings of the 34th Annual Meeting of the Association of Computational Linguistics (ACL'96)*, 310–318, Santa Cruz, California, USA.

Chinchor, N. 1995. "MUC-5} evaluation metrics". In *Proceedings of the Fifth Message Understanding Conference (MUC-5)*, 69–78, Maryland, USA.

Collier, N., H. S. Park, Y. Ogata, Y. Tateishi, C. Nobata, T. Ohta, T. Sekimizu, H. Imai and J. Tsujii. 1998. "The GENIA project: corpus-based knowledge acquisition and information extraction from genome research papers". In *Proceedings of the Annual Meeting of the European chapter of the Association for Computational Linguistics (EACL'99)*. Bergen, Norway.

Craven, M. and J. Kumlien. 1999. "Constructing biological knowledge bases by extracting information from text sources". In *Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology (ISMB-99)*, 77–86, Heidelburg, Germany.

Dempster, A. P., N. M. Laird and D. B. Rubins. 1977. "Maximum likelihood from incomplete data via the EM algorithm". *Journal of the Royal Statistical Society (B)* 39, 1–38.

Fukuda, K., T. Tsunoda, A. Tamura and T. Takagi. 1998. "Toward information extraction: identifying protein names from biological papers". In *Proceedings of the Pacific Symposium on Biocomputing'98 (PSB'98)*, 707–718, Hawaii, USA.

Greenbaum, S. and R. Quirk. 1993. *A Student's Grammar of the English Language*. Essex: Longman.

Kupiec, J. 1992. "Robust part-of-speech tagging using a hidden Markov model". *Computer Speech and Language* 6, 225–242.

Lee, S., N. Collier, T. Ohta, Y. Tateishi, H. Mima and J. Tsujii. 2000. "GPML — GENIA project markup language, version 1.0". Technical report, Tsujii Laboratory, Department of Information Science, University of Tokyo.

McDonald, D. 1996. "Internal and external evidence in the identification and semantic categorization of proper nouns". In Boguraev, B. and J. Pustejovsky (eds.). *Corpus Processing for Lexical Acquisition*, 21–39, Cambridge: The MIT Press.

MUC 1998. *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, Fairfax, VA, USA.

Nobata, C., N. Collier and J. Tsujii. 2000. "Comparison between tagged corpora for the named entity task". In Kilgarriff, A. and T. Berber Sardinha (eds.). *Proceedings of the Association for Computational Linguistics (ACL'2000) Workshop on Comparing Corpora*, 21–27, Hong Kong.

Ohta, T., Y. Tateishi, N. Collier, C. Nobata and J. Tsujii. 1999. "Building an annotated corpus from biological papers". In *Proceedings of the 59th Annual national convention of the IPSJ* (in Japanese), 28–30, Iwate Prefectural University.

Reynar, J. and A. Ratnaparkhi. 1997. "A maximum entropy approach to identifying sentence boundaries". In *Proceedings of the 5th Conference on Applications of Natural Language Processing (ANLP)*, 16–19, Washington DC, USA.

Rindflesch, T., L. Hunter and A. Aronson. 1999. "Mining molecular binding terminology from biomedical text". In *Proceedings of the American Medical Informatics Association (AMIA)'99 Annual Symposium*, 127–131, Washington DC, USA.

Sekimizu, T., H. Park and J. Tsujii. 1998. "Identifying the interaction between genes and gene products based on frequently seen verbs in Medline abstracts". In *Proceedings of the Genome Informatics Workshop*, 62–71, Tokyo, Japan: Universal Academy Press, Inc.

Sekine, S., R. Grishman and H. Shinnou. 1998. "A decision tree method for finding and classifying names in Japanese texts". In *Proceedings of the Sixth Workshop on Very Large Corpora*, Montreal, Canada.

Seymore, K., A. McCallum and R. Rosenfeld. 1999. "Learning hidden Markov structure for information extraction". In *Proceedings of the AAAI'99 Workshop on Machine Learning for Information Extraction*, Orlando, Florida

Tateishi, Y., Y. Ohta, T. Takai and J. Tsujii. 1999. "An ontology for biological reaction events". In *Proceedings of the Tenth Workshop on Genome Informatics*, Tokyo, Japan. Universal Academy Press, Inc.

Tersmette, K., A. Scott, G. Moore, N. Matheson and R. Miller. 1988. "Barrier word method for detecting molecular biology multiple word terms". In Greenes, R. (ed.). *Proceedings of the 12th Annual Symposium on Computer Applications in Medical Care*, 207–211.

Thomas, J., D. Milward, C. Ouzounis, S. Pulman and M. Carrol. 1999. "Automatic extraction of protein interactions from scientific abstracts". In *Proceedings of the Pacific Symposium on Biocomputing'99 (PSB'99)*, 1–12, Hawaii, USA.

Viterbi, A.J. 1967. "Error bounds for convolutions codes and an asymptotically optimum decoding algorithm". *IEEE Transactions on Information Theory*, IT-13(2), 260–269.

Westphal, R., S. Tavalin, J. Lin, N. Alto, I. Fraser, L. Langeberg, M. Sheng and J. Scott. 1999. "Regulation of NMDA receptors by an association phosphatase-kinase signalling complex". *Science*, 93–96.

Wacholder, Y.R., Ravin, Y. and Choi, M. 1997. "Disambiguation of Proper Names in Text". In *Proceedings of the 5th Conference on Applied Natural Language Processing* (ANLP'97), 202–208, Washington DC, USA.

Authors' addresses

Nigel Collier
National Institute of Informatics
2–1–2 Hitotsubashi, Chiyoda-ku, Tokyo 101–8430, Japan
collier@nii.ac.jp

Chikashi Nobata
Keihanna Human Info-Communication Research Center
Communications Research Laboratory
2–2–2 Hikaridai Seika-cho, Soraku-gun, Kyoto 619–0289, Japan
nova@crl.go.jp

Junichi Tsujii
Department of Computer Science
Faculty of Information Science and Technology, The University of Tokyo
7–3–1 Hongo Bunkyo-ku Tokyo 113–0033, Japan
tsujii@is.s.u-tokyo.ac.jp

About the authors

Nigel Collier received his Ph.D. from the Department of Language Engineering at UMIST (UK) in 1996. From 1996 to 1998 he was a Toshiba Fellow researching methods for automatic knowledge acquisition from bilingual corpora. From 1998 to 2000 he worked as a JSPS research associate at the University of Tokyo investigating the application of information extraction to the molecular biology domain. In 2000 he became associate Professor at

the National Institute of Informatics in Tokyo. He currently leads the PIA project at NII developing a portable information extraction system for use on the Semantic Web.

Chikashi Nobata received his Ph.D. from the Department of Computer Science at the University of Tokyo in 2000. The thesis was about the construction of lexical patterns for information extraction. In the same year he joined the Computational Linguistics Group at Communications Research Laboratory, Japan. His current research interests include information extraction and automatic summarization.

Junichi Tsujii received his Ph.D. from the Department of Electrical Engineering, Kyoto University, in 1978. After becoming associate Professor of Kyoto University, Japan and Professor at UMIST in the UK, he has been Professor in the Department of Computer Science, the University of Tokyo since 1995. He is leader of the Information Mobility Project supported by the Japan Science and technology Agency and is also actively involved in the NLP for Biology project. He is a member of the International Committee of Computational Linguistics (ICCL), the liaison officer for Asia of the Association of Computational Linguistics (ACL), and the President of the Association of Natural Language Processing in Japan.