Machine Learning for Information Extraction from XML marked-up text on the Semantic Web

Nigel Collier
National Institute of Informatics (NII)
National Center of Sciences, 2-1-2 Hitotsubashi
Chiyoda-ku, Tokyo 101-8430, Japan
collier@nii.ac.jp

ABSTRACT

The last few years have seen an explosion in the amount of text becoming available on the World Wide Web as online communities of users in diverse domains emerge to share documents and other digital resources. In this paper we explore the issue of how to provide a low-level information extraction tool based on hidden Markov models that can identify and classify terminology based on previously marked-up examples. Such a tool should provide the basis for a domain portable information extraction system, that when combined with search technology can help users to access information more effectively within their document collections than today's information retrieval engines alone. We present results of applying the model in two diverse domains: news and molecular biology and discuss the model and term markup issues that this investigation reveals.

1. INTRODUCTION

The last few years have seen an explosion in the amount of text becoming available on the World Wide Web (Web). Encouraged by the growth of the Web community, textual database providers have been migrating their archives for online access, adding to the available information. Online communities of users have emerged to share documents, and other digital resources in multiple and diverse domains. The issue of information access, i.e. how to find the information that meets users' requirements and present it in an understandable form has now become a major research issue. In order to accomplish this we need to empower computers with an 'understanding' of the users' texts. One scenario that is being explored, e.g. [34], is to combine information retrieval (IR) with information extraction (IE). It is likely that a critical component in such a system will be an ability to learn to identify and classify terms based on examples of previously marked up text.

For this purpose extensible markup language (XML) [35] [14] seems to be most appropriate for semantic annotation, not only for terminology, but also for tasks that require the learning of relations between those terms. XML incorporates a number of powerful features for describing object semantics, such as inheritance of name spaces by child elements from their parents. Name spaces can be nested with the youngest ancestor within a name space declaration determining the current scope. XML allows us to represent semantics through potentially unbounded hyptertags but does not by itself attempt to interpret the meaning of the labels. At the lowest level of IE, a system should be able to identify and classify object, i.e. term, boundaries and classify them according to the semantic classes and ontologies described in the training documents through mechanisms such as the usual document type declaration (DTD) and XML Schemas that are now being proposed [18].

In our work we emphasize the need for IE tools to be adaptable to different domains and languages rather than as general-purpose tools due to the distinct semantic characteristics of each domain.

In the remainder of this paper we present a method for term identification and classification based on hidden Markov models (HMMs) [24] that learns from annotated texts. We describe its performance in two domains, news and molecular-biology and discuss some of the term markup issues that our analysis revealed.

2. BACKGROUND

Information extraction has developed in the last ten years from a collection of ad-hoc methods into a discipline that focuses on a small set of well defined subtasks. This has largely occurred due to the influence of the DARPA-sponsored Message Understanding Conferences (MUCs) in the USA [9][10] and recently the Japanese language IREX conference [29].

In the last few years work has begun on adapting IE for the technical domain of molecular biology, e.g. [6] [8] [16] [26] [27] [31]. As with previous IE approaches, these systems can be classed as either predominantly dictionary-based or learning-based. It is our view that the hand-built dictionary-based systems cannot be expected to be easily ported to new domains and they ignore a potentially valuable source of the domain expert's knowledge, i.e. marked up texts.

Recent studies into the use of supervised learning-based models for the 'named entity' task in both the news and molecular-biology domain have shown that models based on HMMs [1][5][30], maximum entropy [2] and decision trees [28][21] are much more generalisable and adaptable to new classes of words than systems based on traditional hand-

Semantic Web Workshop 2001 Hongkong, China

built patterns and domain specific heuristic rules, e.g. [13], overcoming the problems associated with data sparseness with the help of sophisticated smoothing algorithms [3].

HMMs are one of the most widely methods in ML for IE. They 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 [17]. 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 based on n-grams. 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. Nymble [1], a system which uses HMMs is one of the most successful such systems and trains on a corpus of marked-up text, using only character features in addition to word bigrams.

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 (at the top level of their backoff model) and those which automatically learn about the model's structure such as [30]. Additionally, there is a distinction to be made in the source of the knowledge for estimating transition probabilities between models which are built by hand such as [12] and those which learn from tagged corpora in the same domain such as the model presented in this paper, word lists and corpora in different domains - so-called *distantly*-labeled data [30].

Despite their success, HMMs and other machine learning methods can only be as successful as the features with which they are trained. In this study we have focussed on developing a simple, yet powerful set of features based on orthographic knowledge, that can be portable between domains. We now present an overview of the training corpora we used. This is followed by the results for a HMM named-entity system for two diverse domains: news and molecular biology, using orthographic, lexical and class features to train the models. We also discuss some of the problems that our results revealed, particularly from local syntactic relations, due to the local contextual view the model took.

3. CORPORA

In our experiments we used abstracts in the molecular biology domain available from PubMed's MEDLINE [19] that were marked up in XML by a domain expert [23] as well as a small collection of news texts used in the MUC-6 conference [9] formal and dry runs. It is worth noting that at the present time, no standard test sets exist for the named entity task outside of news (MUC and IREX), making formal system comparisons quite difficult. This will clearly be an important factor in the future development of IE in technical domains.

An example of a marked-up sentence from a news text can be seen in Figure 1. We can see several interesting features of the domain such as the focus of NEs on people and organization profiles. Moreover we see that there are many pre-name clue words such as "Ms." or "Rep." indicating that a Republican politician's name should follow.

In contrast we can see an example of a marked-up sentence for molecular biology in Figure 2. The types of named entities are quite different even on a superficial examination we can see that many are combinations of proper and common nouns with cross-over of vocabulary between name classes, e.g. the lemma 'cell' belongs to both PROTEIN, SOURCE.ct and SOURCE.cl terms.

The sets of name classes for the two domains are given in Tables 1 and 2.

4. METHOD

Our initial approach was 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 considered 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.

The purpose of our model is to find the most likely sequence of name classes (C) for a given sequence of words (W). The set of name classes includes the 'Unk' name class which we use for background words not belonging to any of the interesting name classes given for the two domains (see Tables 1 and 2) and the given sequence of words which we use spans a single sentence. The task is therefore to maximize $Pr(C|W)$. We implement a HMM to estimate this using the Markov assumption that $Pr(C|W)$ can be found from bigrams of name classes.

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 in Section 4.2.

As is common practice, we need to calculate the probabilities for a word sequence for the first word's name class 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,

$$\begin{aligned} Pr(C_t| < W_{first}, F_{first} >) = \\ \sigma_0 f(C_{first}| < W_{first}, F_{first} >) + \\ \sigma_1 f(C_{first}| < \_, F_{first} >) + \\ \sigma_2 f(C_{first}) \end{aligned} \quad (1)$$

and for all other words and their name classes as follows:

$$\begin{aligned} Pr(C_t| < W_t, F_t >, < W_{t-1}, F_{t-1} >, C_{t-1}) = \\ \lambda_0 f(C_t| < W_t, F_t >, < W_{t-1}, F_{t-1} >, C_{t-1}) + \\ \lambda_1 f(C_t| < \_, F_t >, < W_{t-1}, F_{t-1} >, C_{t-1}) + \\ \lambda_2 f(C_t| < W_t, F_t >, < \_, F_{t-1} >, C_{t-1}) + \\ \lambda_3 f(C_t| < \_, F_t >, < \_, F_{t-1} >, C_{t-1}) + \\ \lambda_4 f(C_t|C_{t-1}) + \\ \lambda_5 f(C_t) \end{aligned} \quad (2)$$

where $f(|)$ is calculated with maximum-likelihood estimates from counts on training data, so that for example,

A graduate of <ENAMEX TYPE="ORGANIZATION">Harvard Law School</ENAMEX>, Ms. <ENAMEX TYPE="PERSON">Washington</ENAMEX> worked as a laywer for the corporate finance division of the <ENAMEX TYPE="ORGANIZATION">SEC</ENAMEX> in the late <TIMEX TYPE="DATE">1970s</TIMEX>. She has been a congressional staffer since <TIMEX TYPE= "DATE">1979</TIMEX>. Separately, <ENAMEX TYPE="PERSON">Clinton</ENAMEX> transition officials said that <ENAMEX TYPE="PERSON">Frank Newman</ENAMEX>, 50, vice chairman and chief financial officer of <ENAMEX TYPE="ORGANIZATION">BankAmerica Corp.</ENAMEX>, is expected to be nominated as assistant <ENAMEX TYPE="ORGANIZATION">Treasury</ENAMEX> secretary for domestic finance. Mr. <ENAMEX TYPE="PERSON">Newman</ENAMEX>, who would be giving up a job that pays <ENAMEX TYPE="MONEY">$1 million</ENAMEX> a year, would oversee the <ENAMEX TYPE="ORGANIZATION">Treasury</ENAMEX>'s auctions of government securities as well as banking issues. He would report directly to <ENAMEX TYPE="ORGANIZATION">Treasury</ENAMEX> Secretary-designate <ENAMEX TYPE="PERSON">Lloyd Bentsen</ENAMEX>. Mr. <ENAMEX TYPE="PERSON">Bentsen</ENAMEX>, who headed the <ENAMEX TYPE="ORGANIZATION">Senate Finance Committee</ENAMEX> for the past six years, also is expected to nominate <ENAMEX TYPE="PERSON"> Samuel Sessions</ENAMEX>, the committee's chief tax counsel, to one of the top tax jobs at <ENAMEX TYPE="ORGANIZATION">Treasury</ENAMEX>. As early as today, the <ENAMEX TYPE="PERSON">Clinton</ENAMEX> camp is expected to name five undersecretaries of state and several assistant secretaries.

Figure 1: Example sentences taken from the annotated MUC-6 NE corpus

TI - Activation of <PROTEIN> JAK kinases </PROTEIN> and <PROTEIN>STAT proteins </PROTEIN> by <PROTEIN> interleukin - 2 </PROTEIN> and <PROTEIN> interferon alpha </PROTEIN> , but not the <PROTEIN> T cell antigen receptor </PROTEIN> , in <SOURCE.ct> human T lymphocytes </SOURCE.ct> .
AB - The activation of <PROTEIN> Janus protein tyrosine kinases </PROTEIN> (<PROTEIN> JAKs </PROTEIN>) and <PROTEIN> signal transducer and activator of transcription </PROTEIN> (<PROTEIN> STAT </PROTEIN>) proteins by <PROTEIN> interleukin (IL) - 2 </PROTEIN> , the <PROTEIN> T cell antigen receptor </PROTEIN> (<PROTEIN> TCR </PROTEIN>) and <PROTEIN> interferon (IFN) alpha </PROTEIN> was explored in <SOURCE.ct> human peripheral blood - derived T cells </SOURCE.ct> and the <SOURCE.cl> leukemic T cell line Kit225 </SOURCE.cl> . An <PROTEIN>IL-2</PROTEIN>-induced increase in <PROTEIN>JAK1</PROTEIN> and <PROTEIN>JAK3</PROTEIN>, but not <PROTEIN>JAK2</PROTEIN> or <PROTEIN>Tyk2</PROTEIN>, tyrosine phosphorylation was observed. In contrast, no induction of tyrosine phosphorylation of <PROTEIN>JAKs</PROTEIN> was detected upon stimulation of the <PROTEIN>TCR</PROTEIN>. <PROTEIN>IFN alpha</PROTEIN> induced the tyrosine phosphorylation of <PROTEIN>JAK1</PROTEIN> and <PROTEIN>Tyk2</PROTEIN>, but not <PROTEIN>JAK2</PROTEIN> or <PROTEIN>JAK3</PROTEIN>. <PROTEIN>IFN alpha</PROTEIN> activated <PROTEIN>STAT1</PROTEIN>, <PROTEIN>STAT2</PROTEIN> and <PROTEIN>STAT3</PROTEIN> in <SOURCE.ct>T cells</SOURCE.ct>, but no detectable activation of these <PROTEIN>STATs</PROTEIN> was induced by <PROTEIN>IL-2</PROTEIN>.

Figure 2: Example MEDLINE sentence taken from the XML annotated molecular-biology NE corpus

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*	sublocation
SOURCE.ti	37	*central nervous system*	tissue
UNK	-	*tyrosine phosphorylation*	background words

Table 1: Named entity classes for the molecular biology domain. # indicates the number of tagged terms in the corpus of 100 abstracts.

Class	#	Example	Description
ORGANISATION	1783	*Harvard Law School*	names of organisations
PERSON	838	*Washington*	names of people
LOCATION	390	*Houston*	names of places, countries etc.
DATE	542	*1970s*	date expressions
TIME	3	*midnight*	time expressions
MONEY	423	*$10 million*	money expressions
PERCENT	108	*2.5%*	percentage expressions
UNK	-	*start-up costs*	background words

Table 2: Named entity classes for the news domain. # indicates the number of tagged terms in the corpus of 100 abstracts.

$$f(C_t| < W_t, F_t >, < W_{t-1}, F_{t-1} >, C_{t-1}) \doteq \frac{T(< W_t, F_t >, C_t, < W_{t-1}, F_{t-1} >, C_{t-1})}{T(< W_t, F_t >, < W_{t-1}, F_{t-1} >, C_{t-1})} \quad (3)$$

Where $T()$ has been found from counting the events in the training corpus. In our current system we set the constants λ_i and σ_i by hand and let $\sum \sigma_i = 1.0$, $\sum \lambda_i = 1.0$, $\sigma_0 \geq \sigma_1 \geq \sigma_2$, $\lambda_0 \geq \lambda_1 \ldots \geq \lambda_5$. 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.

In our current system we set the constants λ_i and σ_i 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 [11]. We also impose the restriction that $\sum \sigma_i = 1.0$, $\sum \lambda_i = 1.0$, $\sigma_0 \geq \sigma_1 \geq \sigma_2$, $\lambda_0 \geq \lambda_1 \ldots \geq \lambda_5$. 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 and improve generalisability.

Once the state transition probabilities have been calculated according to Equations 1 and 2, the Viterbi algorithm [33] is used to search the state space of possible name class assignments in linear time to find the highest probability path, i.e. to maximise $Pr(W, C)$.

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%, both for re-tagging spuriously tagged words and for finding untagged words in unknown contexts that had been correctly 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 [25], 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 generalise the HMMs' knowledge about surface forms it is necessary to featurise the vocabulary in some way. On analysing lists of terms we felt that orthographic features offered particularly strong clues about the classes of words, particularly in molecular biology.

Table 3: 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	Examples
TwoDigitNumber	$[25]_{percent}$
FourDigitNumber	$[2000]_{date}$
DigitNumber	$[2]_{percent}$ $[3]_{DNA}$
SingleCap	$[I]_{protein}$ $[B]_{protein}$ $[T]_{source.ct}$
GreekLetter	$[alpha]_{protein}$
CapsAndDigits	$[2A]_{DNA}$ $[BW5147]_{source.cl}$ $[CD4+]_{source.ct}$
TwoCaps	$[RelB]_{protein}$ $[TAR]_{RNA}$ $[HMG]_{DNA}$ $[NF]_{DNA}$ $[NF]_{protein}$
LettersAndDigits	$[p50]_{protein}$ $[Kit225]_{source.cl}$
InitCap	$[Interleukin]_{protein}$ $[Washington]_{person}$
LowCaps	$[kappaB]_{protein}$ $[mRNA]_{RNA}$
Lowercase	$[cytoplasmic]_{source.sl}$ $[tax]_{protein}$
Determiner	the
Conjunction	and
FullStop	.
Comma	,
Hyphen	$[-]_{protein}$ $[-]_{DNA}$
Colon	:
SemiColon	;
OpenParen	(
CloseParen	)
CloseSquare	]
OpenSquare	[
Percent	%
Other	*+#
Backslash	$[/]_{protein}$

Table 3 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 experiments on the 60 news texts using 6-fold cross-validation (50 training, 10 testing) and the 100 molecular biology texts using 5-fold cross-validation (80 training, 20 testing). We then calculated the scores as an average of the F-scores for each marked-up class category.

The results are given as F-scores, a common measurement for accuracy in the MUC conferences that is the harmonic mean of recall and precision. These are calculated using a standard MUC tool [4]. F-score is defined as

System	MUC-6	Biology
HMM with Unity	78.4	75.0
HMM w/o Unity	74.2	73.1

Table 4: Named entity acquisition results for the MUC-6 and molecular biology domains calculated by n-fold cross validation.

$$F - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4)$$

The results are summarised for all classes in each domain in Table 4 and show performance with and without the Unity module. Despite the small number of training texts used, the system could achieve reasonably high performance for both domains. Although results indicate that performance for news texts is slightly better than for biology, this needs confirming with a larger test collection to obtain confidence in the conclusion. The result also highlights the need for soundly motivated metrics (e.g. see [22]) to compare the difficulties of the named entity task between marked-up corpora in different domains. In the following discussion we provide failure analysis of the results.

6. ANALYSIS

In this section we concentrate our discussion on the analysis of results from the molecular-biology corpus. Discussion of term markup issues for the MUC news domain is well documented, e.g. in [20].

During analysis of the corpus we found that a number of syntactic phenomena caused potential complications to identification of term boundaries and their classification. Broadly speaking the major ones 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 knowledge contained within a domain model. The three major issues are now discussed below.

6.1 Coordination

We observed that failure occurred where complex local structures could not be resolved by the shallow-level contextual view of the HMM. 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.

In the following examples, open and close tags have been represented as '[' and ']' respectively for brevity and the tag name appears as a subscript.

Training examples (1), (2) and (3) all 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 (1) and (2) we see that '/' is sometimes used as part of the term and sometimes as a coordinator.

Ex. 1. *...like the [c-rel]$_{protein}$ and [v-rel]$_{protein}$ (proto)-oncogenes.*

Ex. 2. *...regulated by members of the [rel/NF-kappa B family]$_{protein}$*

Ex. 3. *...involves phosphorylation of several members of the [NF-kappa B]$_{protein}$/[I kappa B protein]$_{protein}$ families.*

Examples (2) and (3) resulted in 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 (4) and (5).

Ex. 4. *...regulated by members of the rel/[NF-kappa B family]$_{protein}$*

Ex. 5. *...involves phosphorylation of several members of the [NF-kappa B/I kappa B] protein families.*

Training examples (6) and (7) show more complex cases where the annotation scheme has not allowed the domain expert to fully express her intuition about the classes of terms, particularly where there a sequence of conjoined modifiers. In (6) we see a case of elision where 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 (7) 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.

Ex. 6. *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]$_{protein}$).*

Ex. 7. *.. indicated that multiple regulatory regions including the enhancer, [SP1]$_{protein}$, [TATA]$_{DNA}$ and [TAR]$_{protein}$ regions were important for [HIV]$_{source.vi}$ gene expression.*

It is interesting to note that annotators marking up multi-name expressions in the MUC-6 corpus faced a similar dilemma as noted in [20]. For example "< ENAMEX TYPE= "LOCATION"> North</ENAMEX> and <ENAMEX TYPE-="LOCATION">South America</ENAMEX>. In their case they include the head word with the final modifier as a named entity but make no explicit link to its relation with earlier modifiers.

6.2 Apposition

To quote from [15]: 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 (9), "transcription factor" provides the information that "NF-Kappa B" is a type of protein, and in (8), "RelB-p52" is a "Rel-NF-kappa B complex". i.e. the appositions 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. The challenge posed by appositions is often to know where to start and end the term's boundaries, particularly where no punctuation is used to indicate the apposition phrase's boundary.

Ex. 8. *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]$_{protein}$.*

Ex. 9. *The transcription factor [NF-Kappa B]$_{protein}$ is stored in the [cytoplasm]$_{source.sl}$...*

Apposition by itself in example (8) 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 (10). Interestingly the HMM managed to correctly find the break in the earlier sequence "Rel-NF-kappa B".

Ex. 10. *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 example (9) also posed no difficulty for the HMM as shown in (11).

Ex. 11. *The transcription factor [NF-Kappa B]$_{protein}$ is stored in the [cytoplasm]$_{source.sl}$...*

6.3 Abbreviation

Of more immediate concern to us are cases where abbreviations occurs inside the term itself as shown in the training example (12). Here we require deeper analysis than can be obtained through the local contextual view used by the HMM that we have presented.

Ex. 12. *The [interleukin-2 (IL-2) promoter]$_{DNA}$ consists of several independent [T cell receptor (TcR) responsive elements]$_{DNA}$.*

The result from the HMM for example (12) is given in (13).

Ex. 13. *The [interleukin-2]$_{protein}$ ([IL-2]$_{protein}$) promoter consists of several independent [T cell receptor]$_{protein}$ ([TcR]$_{protein}$) responsive elements.*

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 recognise 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 tags. Although nesting of tags is allowed in XML, cross-over of tags requires a work-around.

7. DISCUSSION

Many of the cases where the model failed to recognise a terms' boundary or class can be considered to be limitations of the local contextual view imposed by the learning model (HMM) and of the mark-up scheme. Although non-nesting of marked-up terms allows us to develop rather simple ML models, it does not reflect the appearance of terms as they are written by domain experts. XML itself does not impose a nesting restriction on us but does insist that there should be no cross-over of name spaces. Although this can be dealt with by workarounds it is not ideal.

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.

For our future work we propose to use shallow dependency analysis to 'normalise' the term, i.e. to disambiguate local syntactic structures. The output of this will be embedded as XML markup into the document for the learning by the ML component of the named entity module.

8. CONCLUSION

In this paper we have presented a scenario for the exploitation of terminology contained within XML marked-up documents that are produced by online domain-based communities for automatically annotating untagged texts based on previously seen examples. Although we have focussed largely on the lowest level of IE, i.e. identification and classification of terms, the idea can be extended to learning higher-level information structures for use in various information access tools as well as long distance dependencies such as anaphora that are necessary to establish equivalence relations between terms, their abbreviations and referencing pronouns.

A limitation of the HMM approach is that it cannot easily model large feature sets due to fragmentation of the probability distribution. In our future work we would like to look at combining orthographic knowledge with other types of lexical information as well as contextual clues from grammatical dependency analysis. It is likely that different ML techniques will be required that can more effectively combine the knowledge from large, possibly sparse, feature sets. The use of Support Vector Machines (SVMs) [32] [7], which have successfully been applied in other research areas, seems one fruitful direction that we are currently exploring.

Acknowledgements

I am grateful to a number of my colleagues at the Tsujii laboratory, University of Tokyo, where most of the experiments for this paper were carried out. In particular I would like to express my gratitude to Sang-Zoo Lee and Chikashi Nobata (CRL) for their helpful comments concerning machine learning methods, Yuka Tateishi and Tomoko Ohta for their efforts to produce the tagged corpus used in the molecular-biology experiments and also to Professor Junichi Tsujii for his support in this work. I would also like to thank the anonymous reviewers for their helpful comments that improved this paper from its draft version.

9. REFERENCES

[1] D. Bikel, S. Miller, R. Schwartz, and R. Wesichedel. Nymble: a high-performance learning name-finder. In *Proceedings of the Fifth Confererence on Applied Natural Language Processing*, pages 194–201, 1997.

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

[3] S. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. *34st Annual Meeting of the Association of Computational Linguistics, California, USA*, 24–27 June 1996.

[4] N. Chinchor. MUC-5 evaluation metrics. In *In Proceedings of the Fifth Message Understanding Conference (MUC-5), Baltimore, Maryland, USA.*, pages 69–78, 1995.

[5] N. Collier, C. Nobata, and J. Tsujii. Extracting the names of genes and gene products with a hidden Markov model. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING'2000), Saarbrucken, Germany*, July 31st–August 4th 2000.

[6] N. Collier, H. Park, N. Ogata, Y. Tateishi, C. Nobata, T. Ohta, T. Sekimizu, H. Imai, and J. Tsujii. 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)*, June 1999.

[7] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, November 1995.

[8] M. Craven and J. Kumlien. Constructing biological knowledge bases by extracting information from text sources. In *Proceedings of the 7th International Conference on Intelligent Systemps for Molecular Biology (ISMB-99)*, Heidelburg, Germany, August 6–10 1999.

[9] DARPA. *Proceedings of the Sixth Message Understanding Conference(MUC-6)*, Columbia, MD, USA, November 1995. Morgan Kaufmann.

[10] DARPA. *Proceedings of the Seventh Message Understanding Conference(MUC-7)*, Fairfax, VA, USA, May 1998.

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

[12] D. Freitag and A. McCallum. Information extraction with HMMs and shrinkage. In *Proceedings of the AAAI'99 Workshop on Machine Learning for Information Extraction*, Orlando, Florida, July 19th 1999.

[13] K. Fukuda, T. Tsunoda, A. Tamura, and T. Takagi. Toward information extraction: identifying protein names from biological papers. In *Proceedings of the Pacific Symposium on Biocomputing'98 (PSB'98)*, January 1998.

[14] I. Graham and L. Quin. *XML Specification Guide*. John Wiley and Sons: New York, 1999; ISBN 0471327530, 1999.

[15] S. Greenbaum and R. Quirk. *A Student's Grammar of the English Language*. Longman, Essex, England, 1990.

[16] K. Humphreys, G. Demetriou, and R. Gaizauskas. Two applications of information extraction to biological science journal articles: Enzyme interactions and protein structures. In *Proceedings of the Pacific Rim Symposium on Bio-Computing 2000 (PSB'2000)*, January 2000.

[17] J. Kupiec. Robust part-of-speech tagging using a hidden markov model. *Computer Speech and Language*, 6:225–242, 1992.

[18] A. Layman, E. Jung, E. Maler, H. S. Thompson, J. Paoli, J. Tigue, N. H. Mikura, and S. De Rose. Xml-data, w3c note 05 jan 1998. The XML-Data specifications are still developing. The latest description can be found at http://www.w3.org/TR/1998/NOTE-XML-data-0105/, January 1998.

[19] MEDLINE. The PubMed database can be found at:, 1999. http://www.ncbi.nlm.nih.gov/PubMed/.

[20] New York University. *Named Entity Task Definition, Version 2.0*, This document can be found online at http://cs.nyu.edu/cs/ faculty/grishman/ NEtask20.book_1.html, May 31st 1995.

[21] C. Nobata, N. Collier, and J. Tsujii. Automatic term identification and classification in biology texts. In *Proceedings of the Natural Language Pacific Rim Symposium (NLPRS'2000)*, November 1999.

[22] C. Nobata, N. Collier, and J. Tsujii. Comparison between tagged corpora for the named entity task. In *Proceedings of the Association for Computational Linguistics (ACL'2000) Workshop on Comparing Corpora, Hong Kong*, October 7th 2000.

[23] T. Ohta, Y. Tateishi, N. Collier, C. Nobata, K. Ibushi, and J. Tsujii. A semantically annotated corpus from MEDLINE abstracts. In *Proceedings of the Tenth Workshop on Genome Informatics*. Universal Academy Press, Inc., 14–15 December 1999.

[24] L. Rabiner and B. Juang. An introduction to hidden Markov models. *IEEE ASSP Magazine*, pages 4–16, January 1986.

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

[26] T. Rindflesch, L. Hunter, and A. Aronson. Mining molecular binding terminology from biomedical text. In *American Medical Informatics Association (AMIA)'99 annual symposium, Washington DC, USA*, 1999.

[27] T. Rindflesch, L. Tanabe, N. Weinstein, and L. Hunter. EDGAR: Extraction of drugs, genes and relations from the biomedical literature. In *Pacific Symposium on Bio-informatics (PSB'2000), Hawai'i, USA*, January 2000.

[28] S. Sekine, R. Grishman, and H. Shinnou. A Decision Tree Method for Finding and Classifying Names in Japanese Texts. In *Proceedings of the Sixth Workshop on Very Large Corpora*, Montreal, Canada, August

1998.

[29] S. Sekine and H. Isahara. IREX: Information retrieval, information extraction contest. In *Information Processing Society of Japan Joint SIG FI and SIG NL Workshop, University of Tokyo, Japan*, http://www.cs.nyu.edu/cs/projects/proteus/irex, September 1998. IPSJ.

[30] K. Seymore, A. McCallum, and R. Rosenfeld. Learning hidden Markove structure for information extraction. In *Proceedings of the AAAI'99 Workshop on Machine Learning for Information Extraction*, Orlando, Florida, July 19th 1999.

[31] J. Thomas, D. Milward, C. Ouzounis, S. Pulman, and M. Carroll. Automatic extraction of protein interactions from scientific abstracts. In *Proceedings of the Pacific Symposium on Biocomputing'99 (PSB'99)*, Hawaii, USA, January 4–9 1999.

[32] V. N. Vapnik. *The Nature of Statistical Learning Theory, 2nd edition*. Springer-Verlag, New York; ISBN 0387987800, 1999.

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

[34] E. Voorhees and D. Harman, editors. *Proceedings of the Eighth Text REtrieval Conference (TREC–8)*, National Institute of Standards and Technology Special Publication, Gaithersburg, Md. 20899, USA, November 17–19 1999.

[35] The XML specifications are developing very rapidly. The latest description can be found in XML 1.0 Recommendation at http://www.w3.org/xml/, 2000.