Bioinformatics Advance Access published July 21, 2005

BIOINFORMATICS

Resolving abbreviations to their senses in Medline

S. Gaudan [a]*; *H. Kirsch* [a], *D. Rebholz-Schuhmann* [a]

[a] *European Bioinformatics Institute, Wellcome Trust Genome Campus,Hinxton, Cambridge, CB10 1SD, UK*

ABSTRACT

Motivation: Biological literature contains many abbreviations with one particular sense in each document. However, most abbreviations do not have a unique sense across the literature. Furthermore, many documents do not contain the long-forms of the abbreviations. Resolving an abbreviation in a document consists of retrieving its sense in use. Abbreviation resolution improves accuracy of document retrieval engines and of information extraction systems.

Results: We combine an automatic analysis of Medline abstracts and linguistic methods to build a dictionary of abbreviation/sense pairs. The dictionary is used for the resolution of abbreviations occurring with their long-forms. Ambiguous global abbreviations are resolved using Support Vector Machines that have been trained on the context of each instance of the abbreviation/sense pairs, previously extracted for the dictionary setup. The system disambiguates abbreviations with a precision of 98.9% for a recall of 98.2% (98.5% accuracy). This performance is superior in comparison to previously reported research work.

Availability: The abbreviation resolution module is available at http://www.ebi.ac.uk/Rebholz/software.html

Contact: gaudan@ebi.ac.uk

1 INTRODUCTION

Abbreviations are a common feature in scientific literature. They are often used without naming the long-form (Fred *et al*. (2004)), resulting in confusion and even in misinterpretations, as soon as the human reader has the wrong long form for the abbreviation on his mind (Sentinel Event Alert (2001)).

We distinguish global abbreviations from local abbreviations. *Global abbreviations* appear in documents without the long-form explicitly stated, while *local abbreviations* come together with their long-form in the document. Global abbreviations are often *ambiguous*, meaning that they have different senses in different documents.

In particular 80% of the abbreviations defined in the Unified Medical Language System (UMLS) have ambiguous occurrences in Medline (Liu *et al*. (2002a)). Regarding human gene symbols from LocusLink, which morphologically are very similar to abbreviations, 40% of the symbols are used in Medline, but many of the occurrences are not related to genes.

Yu *et al*. (2002) also distinguish dynamic and common abbreviations. *Common abbreviations* become accepted as synonyms ("AIDS" and "acquired immunodeficiency syndrome") and represent important terms in their domain, whereas *dynamic abbreviations* are defined for convenience in only a particular article. As a result, global abbreviations are mainly common abbreviations since the reader is expected to know or guess the senses of the global abbreviations.

In the case of local abbreviations the long-form can be retrieved from the document using the extraction method described in Swartz *et al*. (2003)). This improves the precision of gene and protein identification in biomedical text, which suffers from protein/gene symbols that are identical to ambiguous abbreviations. But this method fails in case of global ambiguous abbreviations (Dingare *et al*. (2004)).

Furthermore, many errors in named entity identification are explained by variations observed in the long-forms of abbreviations. For example, AgNor abbreviates two long-forms sharing the same sense: "argyrophilic nucleolar organizer region" and "silver-stained nucleolar organizer region"; similarly, ER abbreviates "estrogen receptor" and "oestrogen-receptor". This property of long-forms is common and has been exploited by Tsuruoka *et al*. (2003) to develop a probabilistic string similarity method.

As a result, resolving local and global abbreviations to their long-forms is a valuable step for improving the quality of Information Extraction and Information Retrieval systems. If the abbreviation is resolved to a normalized long-form, i.e. to a common long-form and not to a minor morphological variant, then this leads to even better results and was persued in our approach.

The most problematic step in abbreviation resolution is retrieving the sense of a global abbreviation that is ambiguous. Stevenson (2002) gives an overview of the state of the art of solving this problem, also known as "Word Sense Disambiguation".

*to whom correspondence should be addressed

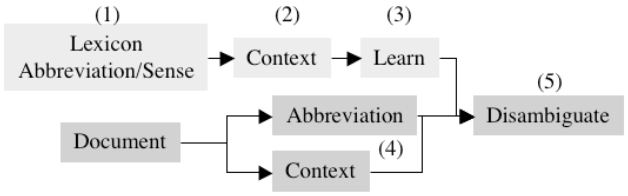

Fig. 1. Disambiguation Process.

The Yarowsky observation (Yarowsky (1995)), which states that terms tend to have "one sense per discourse", provides the foundation for retrieving the sense of a polysemic word by using the context of the document.

Various methods have been implemented for the resolution of ambiguous abbreviations, all following a similar schema (Figure 1). A lexicon is used for collecting the abbreviations and their senses (1). Then the method computes the context of use for each sense (2). Finally, a Machine Learning Algorithm is trained on the context of each sense (3). The disambiguation of an abbreviation contained in a document consists of computing its context in the document (4) and then retrieving the most probable abbreviation sense, given the context (5), thanks to the Machine Learning Algorithm.

This disambiguation schema has been exploited by Pakhomov (2002), Yu *et al*. (2003) and Liu *et al*. (2002b) who use UMLS to collect the abbreviations and their long-forms for the lexicon. However, Pakhomov (2002) observed that not more than one third of the long-forms from UMLS appear in the literature. Furthermore, only frequent abbreviations used in the literature are in UMLS. Since the overlap between the UMLS abbreviations and the ones used in Medline is not sufficient, another dictionary has to be considered. Adar (2004) uses a more relevant approach for the lexicon by using a dictionary extracted from Medline abstracts.

Using the same disambiguation schema, Liu *et al*. (2002b) rely on the UMLS annotations from MetaMap (Aronson (2001)) of the documents as the context of the senses. A Naive Bayes Algorithm is trained on the annotations and then used for the disambiguation, achieving, after removing rare senses, a precision of 92.9% but for a recall of 47.4%.

Concerning the extraction of the context, Adar (2004) relies on the Medical Subject Heading terms (MeSH terms) of the abstracts. The cosine similarity metric is applied for classifying the abbreviation. The method classified correctly 73% of the test set when disregarding rare senses (less than 50 occurrences).

Similarly, Pakhomov (2002) compared a local context[1] with a global context[2] for training a Machine Learning Algorithm based on Maximum Entropy. The system achieves an accuracy of 89% on a limited corpus (10 000 rheumatology notes).

[1] words surrounding the abbreviation

[2] words found in the section containing the abbreviation

Liu *et al*. (2002b) also experimented with the local context and the bag of words technique for training a Support Vector Machine (SVM), reaching an accuracy of 84%.

We present a novel system for the resolution of local and global abbreviations. The resolution of local abbreviations is based on a dictionary of abbreviations, whereas the resolution of global polysemic abbreviations uses a disambiguation process based on the model described in Figure 1.

The first component of the system is a dictionary of abbreviations automatically generated from the literature, inspired by Adar (2004).

Local abbreviations are resolved by looking them up in the dictionary for the most frequent form of the long-form found in the text.

Concerning the resolution of polysemic global abbreviations, we describe first the statistical method used for extracting the context of each sense, and then we explore the disambiguation method based on Support Vector Machines. Finally, we present the global strategy for the resolution of any abbreviation in arbitrary documents.

2 DICTIONARY OF ABBREVIATIONS

The literature is rich in various methods for the automatic extraction of abbreviation/long-form pairs from text. Wren *et al*. (2005) summed up four methods applied to the creation of online databases of abbreviation/long-form pairs when Yoshida *et al*. (2000) focused on the construction of a protein name abbreviation dictionary.

Our abbreviation extraction is based on the method described in Adar (2004), which is robust, fast and achieves a precision of around 95% for a recall of 75%.

An abbreviation is explained in a document by the mention of its long-form. The general pattern is that the long-form is followed by the abbreviation in parentheses; the inverse order of the pair is found at a much lower frequency:

> *The changes in **a**dreno**c**ortico**t**ropin **h**ormone (**ACTH**), cortisol and **de**h**y**dro**e**pi**a**ndrosterone (**DHEA**) in maternal and fetal plasma were estimated in two groups of women.*

After the detection of an abbreviation in parentheses, the correct long-form has to be assigned to the abbreviation. A limited number of rules formalizes how to build abbreviations from a long-form.

The long-form is identified automatically using the Longest Common Subsequence (LCS) in conjunction with a set of scoring rules (Taghva *et al*. (1999)) that favors the first letter of each word of the long form. For each abbreviation candidate (a word surrounded with parentheses), the algorithm matches the long-form in front of the parentheses to the abbreviation and thus determines the boundaries of the long-form.

After scanning all Medline abstracts available in August 2004, the result of our extraction is $5.250.259$ long-form/abbreviation pairs found in $2.857.954$ Medline abstracts. In the following, we refer to this set of abstracts as D.

2.1 Merging Morphologically Similar Long-Forms

Among all extracted long-form/abbreviation pairs, a number of abbreviations share morphologically similar long-forms with the same sense, e.g. "oestrogen receptor" vs. "estrogen-receptor". These long-forms are identified with each other with a similarity measure (Adar (2004)). An n-gram similarity algorithm is used with a cut-off parameter to merge similar long-forms l_1 and l_2:

$$\text{similarity}(l_1, l_2, n) = \frac{|\text{grams}_n(l_1) \cap \text{grams}_n(l_2)|}{\sqrt{|\text{grams}_n(l_1)| \cdot |\text{grams}_n(l_2)|}}$$

$$\text{e.g. grams}_3(\text{'hello'}) = \{\text{'hel'}, \text{'ell'}, \text{'llo'}\}$$

Table 1 illustrates long-forms presenting a high similarity and therefore clustered into groups of long-forms.

The cut-off parameter has been estimated from a hand curated random sample of 250 long-forms doublets. We selected a cut-off parameter of 0.8 so that two long-forms are merged only if they have the same sense.

2.2 Context Based Merging

In contrast to the previous similarity consideration, some long-forms can be morphologically quite different (e.g. "beta site APP-cleaving enzyme" vs. "beta site amyloid precursor protein-cleaving enzyme") but still code for the same meaning. To identify them as synonyms requires domain knowledge, which is provided through the context of the long-forms. Using the context, we merged morphologically diverse long-forms coding for the same meaning.

Adar (2004) relies on the MeSH term annotations of the abstracts for representing the context of the long-forms. However, the granularity of the information contained in MeSH annotations is coarser than the one obtained by extracting relevant words from the text. Furthermore, the MeSH terms approach can not be applied to arbitrary text. As a result, we developed a new method based on word occurrences.

We use here the assumption that two long-forms, coding for the same meaning, are illustrated by documents sharing in average more common words[3] than documents illustrating different meanings (Table 1).

The similarity between two sets of long-forms (g_1 and g_2), created by grouping morphologically similar long-forms together, is computed by considering the number of common words in the sets D_{g_1} and D_{g_2} of documents containing the long-forms, normalized by the total number of words in the documents of the two sets:

$$\text{similarity}(g_1, g_2) = \frac{c(g_1, g_2)}{c(g_1, g_1) + c(g_2, g_2)}$$

with

$$c(g_1, g_2) = \frac{1}{|D_{g_1}| \cdot |D_{g_2}|} \cdot \sum_{d_i \in D_{g_1}, d_j \in D_{g_2}, d_i \neq d_j} \frac{2 \cdot |W(d_i) \cap W(d_j)|}{|W(d_i)| + |W(d_j)|}$$

if $|D_{g_1}|>1$ and $|D_{g_2}|>1$, where $W(d_i)$ is the set of words in the document d_i.

The cut-off parameter has been estimated from a hand curated random sample of 150 long-form set doublets. We selected a cut-off parameter of 0.22 so that two sets of long-forms are merged only if they have the same sense. As a result we use sets (clusters) of long-form/abbreviation pairs which represent the same meaning according to our morphological and contextual similarity estimates. Each cluster contains a number of similar long-forms and the links to the documents containing these long-forms. We define the different senses of an abbreviation by the long-forms found in the different clusters. These abbreviations and their senses are stored in a dictionary.

1) n-gram similarity	
Long-form 1	**Long-form 2**
computed radiography	computed radiographic
compression ratios	compression rate
caloric restriction	calorie-restricted
thrombocytopenia with absent radii	thrombocytopenia and absent radius
transactivator responsive element	trans-activator response element
2) contextual similarity	
Long-form 1	**Long-form 2**
alpha-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid receptors	AMPA receptors
silver-stained nucleolar organizer regions	argyrophilic staining of nucleolar organizer regions
complete remission	complete response

Table 1. Similar long-forms detected with the n-gram similarity (1) and the contextual similarity (2).

3 DISAMBIGUATION OF ABBREVIATIONS

Whenever we find no long-form associated to an ambiguous abbreviation, we use the context to identify the correct meaning of the abbreviation. In the following we describe

[3] Only words included in the pattern (adjective* (proper-noun|noun)$^+$)$^+$ are considered

which suitable context words are generated to disambiguate abbreviations, and how the classifier is trained.

3.1 Context Extraction

The contextual terms used for the disambiguation are extracted using the C-Value algorithm (Frantzi *et al*. (1999)), a method combining linguistical (adjective-noun patterns[4]) and statistical aspects of terms. The C-Value method scores the adjective-noun patterns according to three aspects: the frequency of the adjective-noun patterns (positive correlation), the length of the adjective-noun patterns (positive correlation) and the frequencies of subparts of the adjective-noun patterns (negative correlation):

$$\text{C-value}(w) = \log(\|w\|) \cdot f(w) - \frac{1}{|T_w|} \sum_{v \in T_w} f(v)$$

where

w is the adjective-noun pattern candidate,
$\|w\|$ is the length (in words) of w,
$f(w)$ is the frequency of w in the corpus,
T_w is the set of adjective-noun patterns contained in the candidate w.

Only words contained in terms having a high score are kept for representing a document. After prioritization of the words from the context according to the C-value and applying a cut-off to the list of words, we obtain a tuple Ω=($w_1, .., w_n$) of size n (55 on average) of relevant words for every document.

3.2 The Model

Each abbreviation a belonging to the dictionary has a set of senses, denoted by $S(a)$. Each sense $s \in S(a)$ is illustrated by a set of documents $D_s \subset D$. D_s is the set of documents containing the abbreviation/long-form pairs previously extracted for the construction of the dictionary.

For each document d, the context words are extracted and the document is described by a vector $v = g(d)$ with $g : D \mapsto \{0,1\}^n$. The i^{th} component of v, v_i, is defined as:

$$v_i = \begin{cases} 1 \text{ if the word } w_i \text{ appears in the document } d, \\ 0 \text{ otherwise} \end{cases}$$

As a result, we have a function Φ that associates with each sense s a set of vectors $\Phi(s)$:

$$\Phi(s) = \{g(d) \mid d \in D_s\}$$

3.3 Disambiguation

The task of disambiguating an abbreviation a in a document d is to find the sense $s \in S(a)$ that minimizes the distance between the vector $v = g(d)$ (context of d) and the class defined by $\Phi(s)$.

[4] The following pattern has been used: (adjective* (proper-noun|noun)$^+$)$^+$.

This problem can also be described as a classification problem of assigning $g(d)$ to one of the classes represented by the vector sets $\Phi(s)$ where $s \in S(a)$.

Support Vector Machines (SVM) are suitable classifiers for sparse data in high dimensional spaces and with many relevant features. It has been shown that SVMs achieve substantial improvements over the similar other methods for text categorization (Joachims (1997)). An SVM can separate two classes (positive/negative) by a hyper-plane with a maximum margin between the border vectors. Each class is described by vectors that the SVM "learns". Using the one-against-all approach, we can separate k classes from each other by combining k SVMs. We use in the present case a linear kernel on binary vectors, with an error penalty of 10 in norm L1.

For each sense s of an abbreviation a, we represent the positive class of s by

$$C_+(s) = \Phi(s)$$

and the negative by:

$$C_-(s) = \bigcup_{t \in S(a) \wedge t \neq s} \Phi(t)$$

such that $C_-(s)$ is the set of vectors describing all the senses of a except s. Note that C_- and C_+ are not necessarily disjoint.

A Support Vector Machine is created for each sense s and trained with $C_+(s)$ and $C_-(s)$. The result is a function $h_s : \{0,1\}^n \mapsto \mathbb{R}$ where

$$h_s(g(d)) = \begin{cases} > 0 \text{ predicts } g(d) \in C_+(s) \\ \leq 0 \text{ predicts } g(d) \in C_-(s) \end{cases}$$

For each abbreviation a and for each of its sense, we get the classification functions h_s (a Support Vector Machine). The disambiguation of the abbreviation a in a document d consists of selecting the function h_s so that $h_s(g(d))$ is maximal.

If the resulting $h_s(g(d))$ is positive, then sense$(a, d) = s$ is predicted to be the sense of a in d. If the resulting $h_s(g(d))$ is non-positive, then no sense is predicted:

$$\text{sense}(a, d) = s$$
$$\text{if and only if } \forall \tau \in S(a) : \; h_s(g(d)) \geq h_\tau(g(d))$$
$$\text{and } h_s(g(d)) > 0$$

4 ABBREVIATION RESOLUTION

An aspect of the abbreviation resolution task is the recognition of the abbreviations in the text. Some common English words are also used as abbreviations, making the localization task difficult. The conjunction "if" is used to abbreviate "immunofluorescence" and "for" abbreviates "ferredoxin oxidoreductase". If the document contains sections

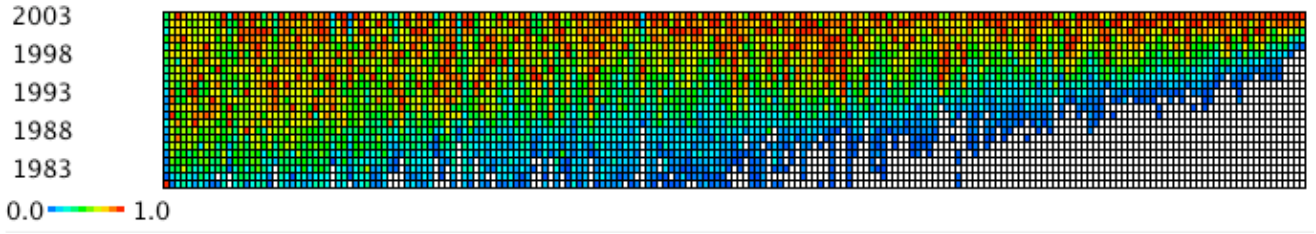

Fig. 2. Abbreviations in Medline over the past 20 years. 200 most frequent abbreviations along the horizontal axis, sorted according to their pattern of occurrence in Medline. The color indicates the relative frequency of the abbreviation. Some of the oldest abbreviations (left part) disappear, reminding of a life cycle. The intensive usage of abbreviations is a recent phenomenon that is in huge progression.

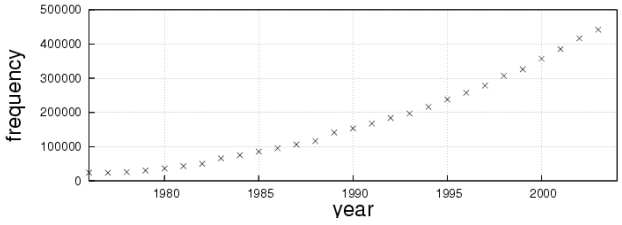

Fig. 3. Number of long-form/abbreviation pairs occurring in Medline since 1975.

	All (1)	**⩾ 20 (2)**	**⩾ 40 (3)**
# Abbreviations	186 641	11 713	7 806
# Ambiguous abbreviations	57 303	3 163	1 851
# Senses	623 441	21 142	12 330
# Occurrences	5 250 259	4 054 993	3 803 758
x̄ long-form variants	1, 57	9, 7	13, 3

Table 2. Counts (#) and averages (x̄) for abbreviation/sense pairs occurring in at least one abstract (1), in at least 20 abstracts (2) and in at least 40 abstracts (3). x̄ long-form variants is the number, in average, of morphological variants per abbreviation's sense.

in uppercase, then the identification task is difficult ("THE" abbreviates "tetrahydrocortisone"). More than 350 abbreviations use the form of a common English word. This problem can be mainly solved by limiting the recognition of abbreviations on adjective-noun patterns, using a Part Of Speech tagger (POS tagger).

When an abbreviation is localized, an efficient search for all the possible long-forms of the abbreviation is applied on the document using a deterministic finite automata (Aho *et al.* (1975)). If a long-form is found, its most frequent form is kept. If no long-form can be retrieved from the document, then a look-up of the abbreviation in the dictionary is performed. If only one sense is found, then the abbreviation is not ambiguous and the most frequent long-form of the unique sense is kept. Finally, if several senses are retrieved, then the disambiguation process is applied.

5 RESULTS

5.1 Dictionary

After mining all Medline abstracts (1965-2004), the dictionary contains 186 641 different abbreviations linked to 623 441 senses, illustrated by 5 250 259 occurrences of an abbreviation with its long-form (Table 2). We distinguished three categories: (1) All abbreviation/long-form pairs, (2) abbreviation/long-form pairs with more than 20 occurrences, and (3) pairs occurring at least 40 times. The third category represents 4% of the total number of abbreviations, but covers more than 72% of the total number of abbreviation/long-form occurrences. We also find in the third category the most morphological variants for the long-forms. As a result, the last category profits the most from normalization of long-forms.

We also found that the number of abbreviations strongly increased over the past 10 years, which correlates the increase of new publications per year. More than half of the abbreviations appeared after 1995 (Figure 3) and last but not least, abbreviation/long-form pairs appear and disappear, similar to the life cycle of gene names in the literature (Hof *et al.* (2003)) (Figure 2). Abbreviation/long-form pairs disappear either because the named concept is not of interest any more or because the abbreviation becomes a common abbreviation, so that the long-form is not provided any more (Figure 4).

The statistics on the dictionary shows that many abbreviations/sense pairs appear at a low frequency (rare abbreviation/sense pairs), whereas few pairs have high frequencies, reminding of the Zipf's law distribution (Figure 5).

The examination of the dictionary shows that some clusters of long-forms should be merged with other ones because their meanings are very similar. But the long-form of these clusters are morphologically very different from each other and their context did not allow to merge the long-forms. However, this phenomenon is only observed on clusters of rare long-forms. For a random sample of 350 long-forms, 42% of the 169 long-forms occurring only once should have been merged with an other entry of the dictionary. This proportion drops to 18% for the 108 long-forms having a frequency

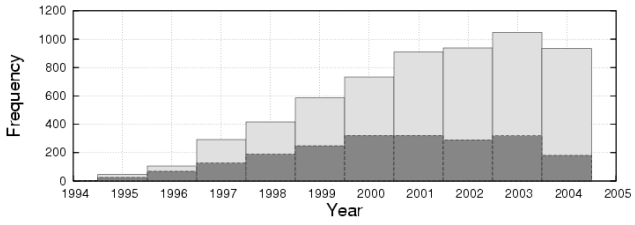

Fig. 4. Frequency of the abbreviation "TUNEL" with its long-form (dark grey) and with or without its long-form (light grey) over the past 10 years (1995-2004). The abbreviation "TUNEL" is not ambiguous in Medline and became a common abbreviation. In 2004, 84% of the occurrences in Medline abstracts of "TUNEL" are without the long-form.

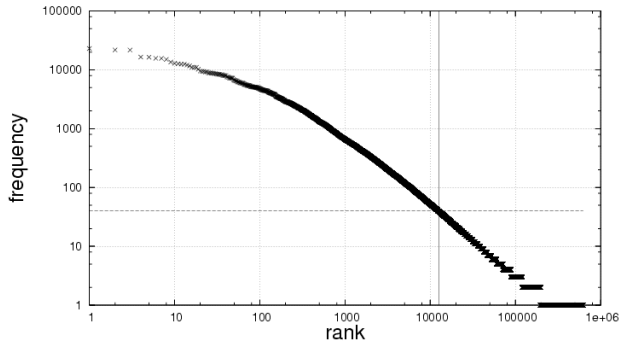

Fig. 5. The rank of the abbreviation/sense pairs and their frequencies, using logarithmic scales. Zipf's law says that there is a constant k such that $frequency(word) \cdot rank(word) = k$. The abbreviation/sense pairs on the left side of the vertical line are pairs occurring at least 40 times.

comprised between 2 and 40. Finally, none of the 73 long-forms occurring at least 40 times share their meaning with other long-forms also occurring at least 40 times.

The dictionary contains many protein and gene symbols that belong to the rare abbreviation sense class (Chen *et al.* (2005)). For example, the gene symbol AFM ("Afamin"), also means "Airflowmeter", "Association Française contre les Myopathies", "acute falciparum malaria", "additive factors method", "aflatoxin M1", "antiferromagnetic" and "atomic force microscopy". Furthermore, in 99% of its occurrences, AFM is used in the sense "atomic force microscopy" and not "Afamin" (Table 3). It is obvious that protein and gene name identification requires the resolution of these symbols.

5.2 Disambiguation

The disambiguation is required for abbreviations having several senses and occurring without the long-form, in other words, for ambiguous global abbreviations. Global abbreviations are also common since they are expected to be known by the reader. As a result, we can apply the disambiguation process using a high quality dictionary by disregarding the rare abbreviation/sense pairs, without changing the nature of the disambiguation problem.

Symbol	HUGO name	#1	Other sense	#2
ACLS	Acrocallosal syndrome	2	advanced cardiac life support	148
ADM	adrenomedullin	254	adriamycin	673
ADMR	adrenomedullin receptor	2	average daily metabolic rate	20
AES	amino-terminal enhancer of split	7	Auger electron spectroscopy	58
AES	amino-terminal enhancer of split	7	anterior ectosylvian sulcus	27
AFA	ankyloblepharon filiforme adnatum	7	amfonelic acid	17

Table 3. List of the 6 first ambiguous abbreviations matching a HUGO symbol and the full name. Column 1 is the HUGO symbol. Column 2 is the full name of the HUGO entry. Column 3 is the number of occurrences found for the corresponding abbreviation. Column 4 is another sense for the abbreviation that occurs more frequently than the full name of the HUGO entry. The column 5 is the frequency of that sense.

In the following we only consider senses which appear frequently enough to profit from disambiguation (40 documents and more). As a result, we have $7\,806$ abbreviations with $12\,330$ senses, representing 72% ($3\,803\,758$) of all pair occurrences in Medline. Out of these $7\,806$ abbreviations, $1\,851$ are polysemic, having in average 3.4 senses with a maximum of 32 senses for "PC" (Table 2).

The Support Vector Machines were trained and tested using a k-fold cross-validation schema (k=5), which measures the quality of predictions on unseen data. For each abbreviation a, a document set is built by grouping the documents illustrating the different senses of a (all the D_s where $s \in S(a)$). Each document set is randomly divided into five subsets equal in size; four are used for the training of the SVM (80%) and one for testing (20%), repeating the operation five times so that each subset has been used for testing. In order to avoid the explicit indication of the sense, the abbreviation long-forms are removed from the text before the SVMs learn or classify the test documents. The system achieves a precision[5] of 98.9% for a recall[6] of 98.2% (98.5% accuracy[7]).

This accuracy can be compared to a baseline derived from a different disambiguation scheme that consists of always selecting the most frequent sense of the abbreviation, independently of the context. Such an algorithm achieves 70% accuracy on the same data.

The accuracy of the disambiguation module has been compared to the disambiguation methods described by Liu *et al.* (2002b), by Yu *et al.* (2003) and by Pakhomov (2002), using

[5] $\text{precision} = \frac{\text{True Positives}}{\text{True Positives + False Positives}}$

[6] $\text{recall} = \frac{\text{TP}}{\text{TP + FN}}$

[7] $\text{accuracy} = \frac{\text{TP + TN}}{\text{TP + FP + TN + FN}}$

the abbreviations used for their tests. Our disambiguation method performs better than their methods for more than 80% of these abbreviations, with an average of 98% accuracy. The remaining 20% are related to abbreviations that have either more or less senses than their test samples.

6 DISCUSSION

The dictionary of abbreviations, the context extraction and the disambiguation module are the three main components of the abbreviation resolution process.

The dictionary has been generated from Medline so that its content is most suitable for abbreviation resolution in biomedical text. The high quality of the dictionary is crucial to achieve the resolution of abbreviations with a high precision/recall. This quality has been reached by combining statistical and linguistical methods for grouping morphological variants of long-forms. Others have also used generated dictionaries, but did not solve the problem of morphological variants for the long-forms or have used external resources (UMLS) that are not suitable when applied on the biomedical literature. The high quality of the abbreviation dictionary has also a direct impact on the accuracy of the disambiguation method. Indeed, the entries of the dictionary are properly linked to the senses of the abbreviations occurring at least 40 times because of the one-to-one relationship between the senses and the entries.

A proper representation of the sense's context is a decisive factor for the discrimination of the senses. We use here a method based on the text itself and not based on human annotations, unlike MeSH terms. Furthermore, the C-Value method provides a refined granularity for the description of the context, without including irrelevant features. The context of a sense is represented with vectors that have on average 3.000 non empty features. In other words, each sense is represented with a considerable number of words.

The accuracy of the disambiguation method profits from the high performances achieved by Support Vector Machines, which have been successfully used in many text classification tasks.

Disambiguation of abbreviations is more accurate than Word Sense Disambiguation on English words because abbreviation's senses are on average more distant. The sense of "tree" as a product of nature and the sense of "tree" as a structure of information are very close. The contexts of both senses can contain "root", "branch", "leaves", even "forest". In contrast, scientific writers tend to avoid to create abbreviations that already exist in their own domain.

Our classifier disambiguates frequent abbreviations in Medline abstracts very accurately. Nevertheless, some misclassifications occur, generally due to one of the following reasons:

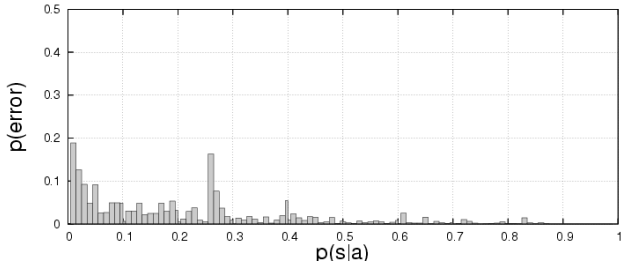

Fig. 6. Distribution of the probability that a misclassification occurs (y axis), given the probability that the the abbreviation take the sense on which the error occurs (x axis).

1. The misclassifications occur for a rare sense of the abbreviation (Figure 6), mainly due to the small margin between scores returned by the positive and negative classification functions (Figure 7). A customized extraction of the context for minor senses could improve the accuracy of the classifier.
2. The misclassification occurs on senses which are very similar but not necessarily synonymous, e.g. "cytotoxic T lymphocyte" and "cytolytic T lymphocyte". These misclassifications can be solved by increasing further the granularity of the context, which becomes difficult to achieve without integrating irrelevant features.
3. The misclassification is due to the fact that some abbreviations are described as ambiguous by the dictionary whereas they are not. According to the dictionary, the abbreviation UDPGT can either take the sense "UDP-glucuronosyltransferase" or the sense "uridine diphosphate glucuronosyltransferase", which are the same. Some further research has to be done for merging these long-forms.

7 CONCLUSION

The biomedical literature contains many abbreviations that can be automatically extracted with their different long-forms. We generated a dictionary of abbreviation/sense pairs, where different morphological variants of a sense have been grouped together with linguistical plus statistical methods.

Using the generated dictionary, local and global abbreviations can be resolved to their sense, using the most frequent long-form as the sense's representative. Many of the extracted abbreviations are ambiguous (1851), meaning that they can take different senses in different contexts. We developed a method that disambiguates the polysemic abbreviations in the documents thanks to the context of it.

On Medline abstracts and for abbreviation/sense pairs which are found at least 40 times our method assigns this sense to the abbreviation with a precision of 98.9% at a recall of 98.2%. The recall is not 100% because, depending on the

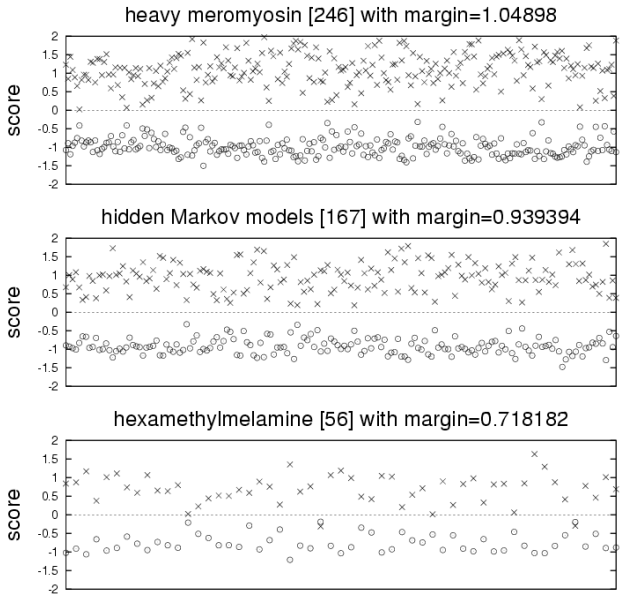

Fig. 7. Scores returned by the SVMs for each sense (class) of "HMM", the positive class (crosses) and the negative class (circles). The distance (margin) between the positive class and the negative class decreases when the senses become rare. The number of abstracts used for the test is given in brackets.

context, the SVM may not assign a sense at all. We assume that abbreviation/sense pairs found less than 40 times are not commonly known and therefore tend to appear with their long-form so that disambiguation is not necessary. There are three reasons for the good performance. First, the senses are, on average, well separated. Second, the method uses a considerable number of relevant words (features) to represent the context of each sense. Third, it has been shown that SVM is the most suitable choice for such data (Joachims (1997)).

Abbreviation resolution can help Information Extraction systems by improving the precision and recall of the recognition of names in documents. Abbreviation resolution can also improve the performances of search engines either by using the resolving abbreviations during the indexing step or by disambiguating the query (query reformulation).

The abbreviation dictionary and the abbreviation resolution module are publicly available.

ACKNOWLEDGMENT

Sylvain Gaudan is supported by an "E-STAR" fellowship funded by the EC's FP6 Marie Curie Host fellowship for Early Stage Research Training under contract number MEST-CT-2004-504640.

REFERENCES

Aho AV., Corasick JM. (1975) Efficient String Matching: An Aid to Bibliographic Search, *CACM*.

Aronson A. (2001) Effective Mapping of Biomedical Text to the UMLS Meta-thesaurus: The MetaMap Program, *Proc. AMIA Symp.2001*, 17-21.

Adar E. (2004) SaRAD: a Simple and Robust Abbreviation Dictionary, *Bioinformatics*, 527-533.

Chen L., Liu H., Friedman C. (2005) Gene name ambiguity of eukaryotic nomenclature., *Bioinformatics*, 248-256.

Dingare S., Finkel J., Manning C., Nissim M., Alex B. (2004) Exploring the Boundaries: Gene and Protein Identification in Biomedical Text, *Proc. of the BioCreative Workshop, Granada*.

Frantzi K., Ananiadou S. (1999) The C value domain independent method for multiword term extraction, *JNLP*, 145-179.

Fred HL., Cheng TO. (2004) Acronymesis: the exploding misuse of acronyms, *Tex Heart Inst J.*, 255-257.

Hoffmann R, Valencia A. (2003) Life Cycles of successful genes, *TRENDS in Genetics*, 79-81.

Joachims T. (1997) Text Categorization with Support Vector Machines: Learning with Many Relevant Features.

Liu H., Aronson AR., Friedman C. (2002) A Study of Abbreviations in MEDLINE Abstracts, *Proc AMIA Symp.*, 464-468.

Liu H., Johnson SB., Friedman C. (2002) Automatic Resolution of Ambiguous Terms Based on Machine Learning and Conceptual Relations in the UMLS, *The Journal of the AMIA*, 621-636.

Pakhomov S. (2002) Semi-Supervised Maximum Entropy Based Approach to Acronym and Abbreviation Normalization in Medical Texts, *Proc. of the 40th Annual Meeting of the ACL*, 160-167.

Sentinel Event Alert (2001) Medication errors related to potentially dangerous abbreviations, *JCAHO*.

Schwartz A., Hearst M. (2003) A Simple Algorithm for Identifying Abbreviation Definitions in Biomedical Text, *Proc. of PSB'03*.

Stevenson M. (2002) Word Sense Disambiguation, The Case for Combinations of Knowledge Sources, *CLSI Studies in Computational Linguistics*.

Taghva K., Gilbreth J. (1999) Recognizing acrronyms and their definitions, *International Journal on Document Analysis and Recognition*, 191-198.

Tsuruoka Y., Tsujii, J. (2003) Probabilistic term variant generator for biomedical terms, *Proc. of the 26th ACM SIGIR*, 167-173.

Weeber M., Schijvenaars B.J.A., van Mulligen EM., Mons B., Jelier R., van der Eijk C., Kors J.A. (2003) Ambiguity of Human Gene Symbols in LocusLink and MEDLINE: Creating an Inventory and a Disambiguation Test Collection, *Proc. of AMIA Annual Symposium*, 704-708.

Wren J.D., Chang J.T., Pustejovsky J., Adar E., Garner H.R., Altman R.B. (2005) Biomedical Term Mapping Databases, *Nucleic Acid Research*.

Yarowsky D. (1995) Unsupervised Word Sense Disambiguation Rivaling Supervised Methods, *Proc. of the 33rd Annual Meeting of the ACL.*, 189-196.

Yu H., Friedman C. (2002) Mapping Abbreviations to Full Forms in Biomedical Articles, *JAMIA.*, 262-272.

Yu Z., Tsuruoka Y., Tsujii J. (2003) Automatic Resolution of Ambiguous Abbreviations in Biomedical Texts using Support Vector Machines and One Sense Per Discourse Hypothesis, *In the Proc. of the SIGIR'03.*, 57-62.

Yoshida M., Fukuda K., Takagi T. (2000) PNAD-CSS: a workbench for constructing a protein name abbreviation dictionary, *Bioinformatics.*, 169-175.