Canadian Artificial Intelligence

Intelligence Artificielle au Canada

Canada's National AI Publication — *La Publication Nationale en IA au Canada*

Winter 2000 — *No. 45* — *hiver 2000*

Contents / Contenu

Canadian Artificial Intelligence is published by the Canadian Society for Computational Studies of Intelligence (CSCSI). ***Intelligence Artificielle an Canada*** est publiée par la Société canadienne pour l'étude de l'intelligence par ordinateur (SCEIO). Canadian Publications Mail Product Sales Agreement No. 1497200.

ISSN 0823-9339

Web Version Web

Please **don't throw away** the **envelope** in which CAI/IAC arrived. It contains information about your userID and password, which are needed to access the members-only area of the CSCSI/SCEIO website:

http://cscsi.sfu.ca/cai.html.

Sample issues and articles are accessible to non-members. The members-only area contains this issue (#45) and some past issues of *CAI/IAC*. To access the area, type your userID and password at the login window. Your userID is the first letter of your first name plus up to seven letters of your last name.

For example, the userID for Anne Murray is amurray. Your password is based on your CSCSI/SCEIO membership number which is **printed on the envelope** in which this issue arrived. Take that number and prepend to it the first letter of your first and last name. (e.g. if Anne Murray's membership number was 876543, then her password would be am876543.)

President's Message

Robert Mercer

CALL FOR NOMINATIONS

CSCSI Distinguised Service Award

This award is presented biannually to an individual who has made outstanding contributions to the Canadian AI community in one or more of the following areas:

- Community service
- Research
- Training of students
- Research/Industry interaction

The award, which will be presented at the CSCSI conference, provides:

- honorary lifetime membership in CSCSI
- conference fees when the award is presented
- a token gift

Recommendations for the award should be addressed to me to arrive by **31 March 2000**. They should include a brief (1 page) summary of the nominee's qualifications for receiving the award. The final decision will be made by the CSCSI executive.

It is very much hoped that the winner will be present to receive the award in person, and we ask nominators to make discreet enquiries in advance to ensure that there is a reasonable likelihood of their nominee attending the conference.

APPEL DE CANDIDATURES

Prix de distinction de la SCEIO pour service rendu

Cette récompense est présentée deux fois par année à un individu qui aura fait une contribution hors de l'ordinaire à la communauteé d'Intelligence Artificielle au Canada dans au moins un des secteurs suivants:

- Service à la communauté
- Rescherce
- Formation d'étudiants
- Intéraction Recherche/Industrie

La récompense qui sera présentée à la confèrence de la SCEIO, comprend:

- Carte de membre à vie à la SCEIO à titre de membre honoraire
- Accès gratuit à la conferénce à laquelle la récompense sera attribuée
- Cadeau d'appréciation

Les nominations devraient m'être adressé personnellement au plus tard le **31 Mars 2000**. Elles devraient inclure un bref (1 page) sommaire des qualifications du candidat. La décision finale sera rendu par l'exécutive de la SCEIO.

Il serait souhaitable que le gagnant puisse venir chercher sa récompense en personne, et nous demandons aux parrains de s'informer auparavant de la disponibilité de leur candidat, quant à la possibilité que le candidat assiste à la conférence en vue de recevoir son prix.

Nominations for CSCSI Distinguished Service Award are to be addressed to:

Robert Mercer
Department of Computer Science,
University of Western Ontario,
London, ON, N6A 5B7
email: mercer@csd.uwo.ca

NEWS
NOUVELLES

NSERC $1 Million Research Prize

In honour of Nobel Prize Winner Gerhard Herzberg (Montréal, Québec) - Secretary of State (Science, Research and Development) Dr. Gilbert Normand has announced a $1 million science and engineering research prize in honour of the late Gerhard Herzberg, winner of Canada's first Nobel Prize for research in Chemistry. The annual award is a millennium project of NSERC (the Natural Sciences and Engineering Research Council of Canada).

In announcing the prize, Dr. Normand said, "This dramatic new award, which guarantees the recipient $1 million in research funding, will celebrate great Canadian achievements in research and move the best Canadian researchers to a new level of research support. It will inspire all Canadians, and especially young citizens, with the excitement of science and engineering in Canada."

The inaugural competition for the Herzberg Medal will be held in 2000. The winner will be selected from three finalists. The others will receive a one-time award of $50,000 of research support.

The new prize guarantees that the winner will have a million dollars for his or her own research or to direct the use of in some related way (e.g. scholarships or the establishment of a university chair)

If the winner already has an NSERC Research Grant, his or her grant for each of the five years will be brought up to $200,000. If the grant is greater than $150,000, it will be topped up by $50,000.

A winner who is not an NSERC grant recipient may direct the full $200,000 to scholarships or the establishment of a chair.

Nomination kits for the Gerhard Herzberg Canada Gold Medal for Science and Engineering will be available from NSERC in early January and will also be available on the Web at www.nserc.ca/about/awardeng/htm

NSERC is the primary federal agency investing in people, discovery, and innovation. The Council supports both basic university research through research grants and project research through partnerships of universities with industry. NSERC also supports the advanced training of highly qualified people in both areas.

Source: http://www.nserc.ca/news/p991125.htm

Audio Signal Classification: An Overview

David Gerhard

Résumé

La classification de signaux audio consiste a extraire les caractéristiques physiques et perceptuels d'un son, et d'utiliser ces catéristiques pour identifier l''ensemble de classes d'appartenance le plus approprié pour le son en question. Les algorithmes utilisés dans l'extraction et la classification de caractéristiques peuvent être plutôt variés selon le domaine de classification de l'application. Cet article présente une vue d'ensemble de l'état de la littéreture actuelle dans le domaine de la recherche en classification de signaux audio.

Abstract

Audio signal classification consists of extracting physical and perceptual features from a sound, and of using these features to identify into which of a set of classes the sound is most likely to fit. The feature extraction and classification algorithms used can be quite diverse, depending on the classification domain of the application. This paper presents an overview of the current state of the audio signal classification research literature.

1 Introduction

Audio signal classification (ASC) is a field of research that, historically, has been explored in a few very concentrated areas, with less work done on the general problem. The individual pieces have traditionally been speech recognition and related problems, music transcription [Moo77] [Pis86], and recently speech/music discrimination [Sau96] [SS97]. Other problems in the field have been researched as well, but the main direction has been toward speech and music applications.

The major exception that has recently appeared is the multimedia database [FG94] [WBKW96]. Searching algorithms that find a piece of multimedia based on the sound must be robust and must consider much more than simply music or speech, and for that reason an algorithm that sorts sounds into appropriate categories must be able to deal with a much larger range of sounds that a speech recognizer would. This is not to say that the problem is any harder — a speech recognizer only expects speech, but must distinguish between all possible phonemes in the applicable language, as well as recognizing slurs between phonemes, word boundaries, and sometimes must even glean meaning from prosodic features such as pitch, emphasis and timing.

1.1 Classification: Problems, Applications

Speech recognition is perhaps the classic ASC problem [RJ93]. Incoming sound must be classified by the phoneme being voiced at any particular time, and then the phonemes must be combined to form the most likely words, phrases and sentences that make sense. Speech recognition is only recently beginning to be a useful tool instead of a frustrating exerc13ise, but it is still a long way from the dream of speaker-independent continuous recognition of non-grammatical speech in a noisy environment. There is much ASC work to be done.

Music recognition and transcription is another computing application of ASC. Here, The sound to be classified is music, the possible classes are musical notes, and the output is usually a musical score. Systems have been developed to solve this problem with one or two musical lines [Moo77], but when the sound to be transcribed contains a full orchestra, the problem is considerably harder. Audio signal classification would speed up the creation and use of melody databases which could be accessed by direct human input, in the form of humming, whistling or singing at the computer.

A more general ASC system could differentiate between speech and music, and could be used to assign a sound to a particular transcription system. If a sound were classified as speech, the speech recognizer would be employed, and if were music, the music transcriber would be called into service. In a system such as this, the speech recognizer and music transcriber could be optimized to expect only the appropriate input, simplifying each and improving the robustness of the whole system.

More publicly consumable applications of ASC include television and radio advertisement identification, for muting or VCR pausing, and radio music style recognition that would receive a command from a user, such as "Play me some country!" and scan the available radio channels for music fitting the query. A telephone on-hold indicator could be developed to recognize when the piped-in music has stopped and a person is on the line.

ASC could be applied to equalization, the boosting or attenuating of particular frequencies in a sound, to make it more intelligible or enjoyable. Because this process has traditionally been done by hand, a finite number of equalization filter settings are often employed in an audio application. Equalization filters are currently used in hearing aids as well as in public address systems, where the user must change the setting by hand, depending on the input. An automatic equalization system could detect the current environment and apply an appropriate filter setting.

Each of these possible applications represents a separate ASC problem which must be examined and researched individually. The quest for the universal classifier is probably as futile in the audio field as it is in any other classification field at this point.

2 Features

The first step in any classification problem is to identify the features that will be used to classify the data. Feature extraction is a form of data reduction, and the choice of feature set can make or break a classification system. For that reason it is sometimes left to the brute power of an algorithm to decide which features are the most important. For more on automatic selection of features, see [Sch96] [DH73].

2.1 Physical and Perceptual Features

The features typically used in ASC can be divided into physical and perceptual categories. *Physical features* are properties that correspond to physical quantities, such as fundamental frequency (F_0), energy, zero-crossing rate (ZCR) and modulation rate. *Perceptual features* are properties that correspond to the way humans perceive sound. These include pitch, loudness, timbre and rhythm. Some features are related, such as pitch with F_0 and energy with loudness, but it is important to recognize that these features are not identical. Pitch is related to the log of frequency, but pure sinusoids sound sharp, and very low frequencies sound flat, compared to a purely logarithmic relationship [CWE94]. For a feature to be truly perceptual, there must be some perceptual model in the extractor used to measure the feature.

As this paper is meant to be an overview, only two features will be presented in detail along with the more general discussion of feature extraction and analysis.

2.2 Fundamental Frequency

The *fundamental frequency*, or F_0, of a signal is the lowest frequency at which the signal repeats, and is only relevant for periodic or pseudo-periodic signals. It is clear that extracting the F_0 from a signal will only make sense if the signal is periodic. Because of this, F_0 detectors can be used as periodicity detectors — if the F_0 extracted makes sense for the rest of the signal, then the signal is considered to be periodic. If the F_0 appears to be randomly varying or is detected as zero, then the signal is considered to be non-periodic.

F_0 can be used as a classification feature in many ways. If the F_0 changes in discrete jumps and stays on particular values, it is a good indicator of music. If a F_0 pattern repeats regularly but within the pattern the F_0 tends to sweep across values, the sound could be a birdcall. Most of the information that is obtained from examining the F_0 of a sound comes from watching how the sound changes over time.

2.3 Zero Crossing Rate (ZCR)

As the name implies, ZCR is a measure of how often the sound signal crosses from positive to negative or vice-versa. At first investigation, it might seem that ZCR would be a good technique for finding F_0, the thought being that a sinusoid will cross the zero line twice per cycle and hence ZCR should be exactly $2F_0$. It was soon made clear that there are problems with this measure of F_0 [Roa96]. If the signal is spectrally rich, then it might cross the zero line more than twice per cycle. It also became clear that ZCR is useful for extracting features other than F_0, such as spectral content, voicedness and noise content [Ked86]. Many recent classification systems employ various statistics taken from the ZCR measure [SS97] [Sau96]. ZCR is a popular feature because it is very easy to measure — analog ZCR meters detect the change in polarity of the voltage from a sound input line without a single cycle of digital signal processing..

2.4 Feature Duration

It is often important to examine how the feature values change over time. The instantaneous F_0 of a signal is instructive, but how the F_0 changes is often much more useful. It can be used to investigate the constancy of the F_0, and the duration of each sound event that has frequency, as well as how much of the total signal is taken up with periodic sound. These sub-features can be applied to many other features. For example, a common feature used in speech classification systems is whether the sound is voiced or unvoiced. If the sound is periodic, it is most likely voiced, and if it is non-periodic with strong high-frequency components, it is likely to be unvoiced. The ratio between voiced and unvoiced segments in the sound can be used in a speech/song decision, because a piece of singing is more likely to have long chunks of voiced sound punctuated by short breaks of unvoiced sound.

Another property based on feature duration is the perceptual property of rhythm. When a piece of sound is considered rhythmic, it often means that it repeats in some way, on a time scale much longer than that required to generate a frequency. Periodicity in feature values is a good indicator of rhythmicity, and this requires examining the sound at a much longer scale than most feature extractors use. Since rhythmic information, along with other types of information, can occur at many different scales, it is often useful to examine the signal at more than one resolution at a time. Techniques that do this are called *multiresolution* techniques.

2.5 Feature Clustering

Some classifications can be determined from a single feature, but most are confirmed by examining several features at once. Algorithms that do this statistically, called clustering algorithms, make use of many pieces of data. Each piece of data, called a *case* in the clustering literature, corresponds to an observation of a sound, and the features extracted from that observation are called *parameters*. Clustering algorithms work by examining a large number of cases and finding groups of cases with similar parameters. These groups are called *clusters*, and are considered to belong to the same category in the classification. Once the clusters have been discovered, a representative case is chosen for each cluster, usually corresponding to the center of each cluster, and new cases are classified depending on the proximity to the representative cases. More detailed discussions of clustering techniques are presented in [Sch96] and [DH73].

3 Conclusion

Pattern Classification is a traditional and well-studied problem, and ASC is a specific instance of that problem which has not been studied in as much detail. In itself, ASC is a broad research area which can be divided into many sub-problems. ASC has many potential applications and many open research problems.

References

[CWE94] Stanley Coren, Lawrence M. Ward, and James T. Enns. *Sensation and Perception*. Harcourt Brace College Publishers, Toronto, 1994.

[DH73] Richard O. Duda and Peter E. Hart. *Pattern Classification and Scene Analysis*. John Wiley and Sons, Toronto, 1973.

[FG94] Bernhard Feiten and Stefan Günzel. Automatic indexing of a sound database using self-organizing neural nets. *Computer Music Journal*, 18(3):53–65, Fall 1994.

[Ked86] Benjamin Kedem. Spectral analysis and discrimination by zero-crossings. *Proceedings of the IEEE*, 74(11):1477–1493, November 1986.

[Moo77] James A. Moorer. On the transcription of musical sound by computer. *Computer Music Journal*, pages 32–38, November 1977.

[Pis86] Martin Piszczalski. *A Computational Model for Music Transcription*. Ph.D. thesis, University of Stanford, 1986.

[RJ93] Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of Speech Recognition*. Prentice Hall, New Jersey, 1993.

[Roa96] Curtis Roads. *The Computer Music Tutorial*. MIT Press, Cambridge, 1996.

[Sau96] John Saunders. Real-time discrimination of broadcast speech/music. In *International Conference on Acoustics, Speech and Signal Processing*, pages 993–996. IEEE, 1996.

[Sch96] Jürgen Schürmann. *Pattern Classification*. John Wiley and Sons, Toronto, 1996.

[SS97] Eric Scheirer and Malcolm Slaney. Construction and evaluation of a robust multifeature speech/ music discriminator. In *International Conference on Acoustics, Speech and Signal Processing*, volume II, pages 1331–1334. IEEE, 1997.

[WBKW96] Erling Wold, Thom Blum, Douglas Keislar, and James Wheaton. Content-based classification, search and retrieval of audio. *IEEE MultiMedia*, pages 27–37, Fall 1996.

About the Author

David Gerhard (http://www.cs.sfu.ca/~dbg) is a PhD Student in the School of Computing Science, Simon Fraser University, Burnaby, BC V5A 1S6, Canada. His research interests include Sound Classification, Audio Visualization and Auditory Scene Analysis.

Using Data Mining for Crop Genebank Management

Raju Addala, Ken Whelan, Janice Glasgow

Résumé

Cet article décrit la continuation de notre travail dans l'application de techniques de machines apprantissantes pour optimiser la gestion de moissons de banques de gènes. En particulier, nous montrons comment nous pouvons identifier les gènes souhaitables (gènes résistants aux maladies de moisson) basès sur de petits sous-ensembles (collections de noyau) de la collection de large/base. Une vue d'ensemble des moissons de banques de gènes et des collections de noyau est fournie. Nous abordons quelques questions concernant la formation et l'utilisation de collection de noyau à partir d'une perspective de machine apprantissante. Enfin nous fournissons un exemple pour illustrer nos idées.

Abstract

This paper describes our ongoing work on the application of machine learning techniques to optimize the management of crop genebanks. In particular, we show how we can identify desirable genes (genes resistant to crop diseases) based on small sub-sets (core collections) of the large/base collection. An overview of crop genebanks and core collections is provided. We address some issues in forming and using core collections from a machine learning perspective. Finally we provide an example to illustrate our ideas.

1 Introduction

Worldwide efforts to conserve crop genetic resources (CGR) have resulted in large crop genebanks. Currently, over six million accessions of various crop species are preserved in genebanks. Paradoxically, the use and management of the genebanks is hindered by the very success in collecting such a vast resource. To mitigate this problem, the core collection concept was proposed [11, 5]. A core collection is a sub-set (approximately 5-10%) of the base collection. The concept is based on principles from statistical and population genetics [3]. The underlying idea is that the information gained from the core collection can be used to extract useful genes from the base collection. However, no scientific studies have yet been initiated to extract and use the information from the core collections. The objectives of this paper are: (a) to provide an overview of the crop genebank and core collections, and (b) to discuss how data mining can be used to manage genebanks. The data mining techniques used for the second task include a number of well-known machine learning algorithms that discover useful knowledge from large databases. The discovery process most relevant to genebanks is the representation and extraction of knowledge from core collections that can then be used to identify useful genes/alleles from the genebanks.

2 Crop Genetic Resources and Their Importance

Crop genetic resources (CGR) — or germplasm — are seeds, plants, or plant parts that are useful in crop breeding, research, or conservation because of their genetic attributes [7]. The importance of conserving these resources has long been realised and now there are approximately 6 million accessions preserved in the genebanks [10]. To overcome the problems posed by the collections' size, Harlan [14] recommended that a small subset should be carefully sampled to represent the diversity of the base collection. This concept was termed the core collection by Frankel [11]. Formally, a *core collection* is defined as a "sub-set derived from an existing germplasm collection, chosen to represent the genetic spectrum in the entire collection [11,]". The primary motivation behind the concept is to save resources. Ideally, ***a core collection should act as a window to the base collection***, have no redundant entries, and useful genes should be readily found. It should also have high genetic diversity or allelic richness.

2.1 Open issues in Core Collections

Despite the wide acceptance of the core collection concept, two central issues remain unresolved. These are: (a) What is the best strategy for developing core collections? and (b) How to extract, formalize and use information from core collections to predict/identify useful genes/alleles from the base collection? Table 1 provides an overview of the various methods for developing core collections. Table 2 provides an evaluation of the various sampling methods and shows that not all sampling methods are tested on a single crop species (with exception of [1]) and none of the studies have compared the predictive accuracy of the sampling methods.

	Sampling strategy	Reference	Prerequisite		
			Passport data	Evaluation data	Genetic data
Group A	Random (R)	[4]	No	no	no
	Sequential (S)		No	no	no
	Equidistant (E)		No	no	no
Group B	Constant (C)	[4]	yes	no	no
	Proportional (P)	[4]	yes	no	no
	Logarithmic (L)	[4]	yes	no	no
Group C	Principal components score strategy (PCSS)	[13]	no	yes	no
Group D	Hierarchical cluster-based strategies	[16]	Yes/no	yes	no/yes
	Stratified canonical strategy (CAN)	[21]	yes	yes	no
Group E	H-strategy (H)	[18]	yes	no	yes
	M-strategy (M)	[18]	yes	no	yes

Table 1: Sampling methods

	Size		Evaluation criteria						
Sampling methods compared	Base	Core	Passport	No.of alleles	No. of desirable alleles	Genetic background	Genetic diversity	Predictive accuracy	Source
R,C,P,L	1358	140	yes	yes	no	no	yes	no	[3]
R,C,L,P,M	2500	250	yes	yes	no	no	no	no	[2]
R,S,C	2000	200	yes	yes	no	no	no	no	[20]
R,C,P,L,Cl	96	10	no	yes	no	no	no	no	[19]
R,C,P,L,CAN	3000	500	no	yes	no	no	yes	no	[21]
R,E,P	105	10	yes	no	no	no	no	no	[9]
R,C,P,L	1240	211	yes	no	no	no	no	no	[8]
R,E,C,P									
L,CAN,PCA									
PCSS	6300	630	yes	yes	yes	yes	yes	no	[1]

Table 2: A review of comparing the efficiency of the sampling methods

2.2 Core Collection: A Machine Learning Perspective

Machine Learning is a field concerned with developing techniques to automatically extract knowledge from collected information presented as examples. In supervised learning this knowledge takes the form of a *classifier.* Typically the knowledge is extracted from a subset (the training examples) of the collected information and the performance of the resulting classifier is measured on a different subset (the test examples). A current research issue in Machine Learning is the efficient extraction of knowledge from very large databases, a similar problem to the construction of core collections from genebanks.

These problems are both addressed by sampling a small but relevant subset of the data. Questions central to generating a most relevant sample are: 1) how to select the training examples comprising the sample (what is the best sampling method?) and 2) what is the minimum number of training examples necessary to maintain the best performance of the learning algorithm. In this context, the concept of a sample is comparable to the concept of a core collection. By applying machine learning in this way, a core collection may indeed become a more reliable window to the base collection. This is done by uncovering knowledge in the core collection which may be used to predict useful genes in the base collection.

The identification of the best sample subset is a combinatorially expensive operation. Formally, there are 2^N possible sub-sets, where N is the size of the base collection. The crux of the problem is to find the smallest "n" (size of the core collection) that generates the required knowledge to successfully predict and identify desirable genes from the genebank.

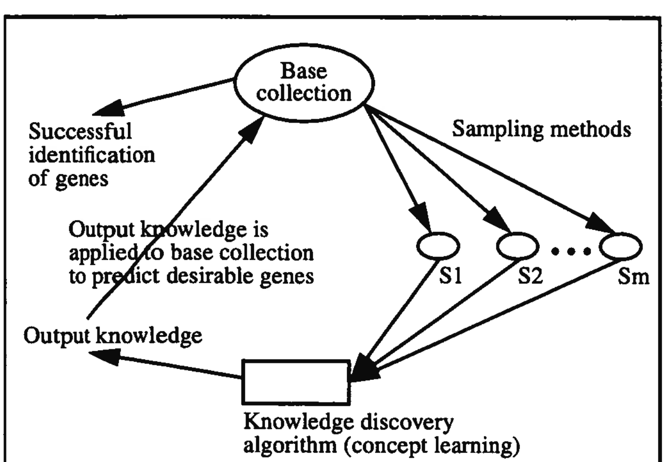

Figure 1: Methodology for applying Data Mining to Genbank Management.

3 Experimental Procedure

3.1 Forming the Core Collections

Sub-sets (core collections) are formed using the sampling strategies in Table 1. A number of machine learning techniques are used to construct classifiers from each of the subsets. These classifiers represent relationships between the attributes used to describe features of the plants and the attributes used to describe the genes responsible for stresses, e.g. resistance to an insect pest. The classifiers can then be applied to the base collection to predict the desirable genes. In addition, the comprehensibility of the discovered knowledge can also be investigated (another important issue in data mining).

3.2 Determining the Smallest Sample Size

For each of the above sampling methods, we are applying the "progressive sampling method" to determine the smallest sample size that maintains the maximum predictive accuracy [17]. *Progressive sampling* starts with a small sample and uses progressively larger ones until the predictive accuracy no longer improves. A central component of progressive sampling is a sampling schedule $S = \{n_0, n_1, .., n_k\}$ where each n_i is an integer that specifies the size of a sample provided to the learning algorithm ($n_i \leq N$ for all i).

3.3 Evaluating the Sampling Methods

Since the purpose of a core collection is to enhance the use of CGR, a good core collection should not only provide a high frequency of desirable genes/alleles but it should also be able to identify additional desirable genes from the base collection with no extra effort. Thus the frequency of rare and/or desirable characters and the predictive power may be used to judge the merit of a core collection. The relative efficiency of various sampling strategies for developing core collections has been compared by different researchers. However, none considered or compared the predictive power of the sampling methods (see Table 2). We propose to determine the best sampling method as the one that produces the smallest core collection that allows the generation of the most predictive classifier with the minimum classification error.

4 A Case Study from ICARDA Chickpea Genebank

Chickpea is a self pollinated diploid (2N=2x=16) of the mongeneric tribe Cicereae Alef. (Leguminosae) and is the third most important food legume in the world. This crop is receiving increased attention in Australia and

North America, especially Western Canada. The ICARDA (International Center for Agricultural Research in the Dry Area, Aleppo, Syria) chickpea collection (ICARDA collection) consists of 9586 accessions of cultivated chickpea and 291 accessions of wild chickpea [15]. In 1987, 6330 accessions, consisting of 5163 landraces and 1167 breeding lines from 35 countries, were evaluated for 24 characters at Tel Hadya, Syria during the winter of 1987-88. Ghab 1 and Ghab 2 were planted as checks. Twenty-two of these 24 characters (13 morpho-physiological characters, three phenological characters, and response to six stresses) will be used to compare the sampling methods for developing core collections (Table 3). Sixteen of these 22 attributes will be used to predict the goal/target attributes (six attributes).

4.1 An Illustrative Example

The chickpea genebank is used here to illustrate how machine learning can be used to identify/predict desirable genes from genebanks. The task in this case is to identify plants resistant to leaf miner (an insect pest effecting the chickpea plant). 1600 plants were sampled from the base collection and the population was divided into two equal sets (training set and test set). Two machine learning algorithms were used to generate classifiers from the training set. The machine learning techniques used were CN2 and a neural network. The results from the experiments performed with each technique are described below.

Type of character	Name of character (range)	Domain
Morpho-physiological	Growth habit (3)	erect, semi-erect, spread
	Canopy width (20-96 cm)	continuous
	Seeds/m2 (3-2467)	continuous
	100-seed weight (8.4-70.1 g)	continuous
	Seed type (3)	kabuli, desi, pea
	Seed coat colour (8)	beige,yellow,white,orange light-brown,brown,green
	Seed coat texture (3)	rough, smooth, tuber
	Plant height (25-85 cm)	continuous
	Pod dehiscence (3)	1,2,3
	Biological yield (25-1080 g/ plant)	continuous
	Seed yield (1-510 g/plant)	continuous
	Harvest index (1-78 %)	continuous
	Protein content (13.5-28.2%)	continuous
Phenological	Days to flowering (115-156)	continuous
	Flowering duration (12-83)	continuous
	Days to maturity (174-206)	continuous
Stress response (Goal attribute)	Ascochyta-blight resistance (2)	resistant, susceptible
	Cyst-nematode resistance (2)	resistant, susceptible
	Leaf-miner resistance (2)	resistant, susceptible
	Cold tolerance (2)	tolerant, susceptible
	Herbicide tolerance (2)	tolerant,susceptible
	Iron chlorosis (2)	tolerant, susceptible

Table 3: Knowledge base for ICARDA chickpea collection.

4.1.1 Neural Networks

We used the backpropagation algorithm implemented by Goodman [12]. The network topology consisted of 16 input nodes, 16 hidden nodes and one output node (leaf miner resistance or susceptible). All the input nodes were connected to all the hidden nodes and all the hidden nodes were connected to the output node. In addition input node 1 (passport data/country of origin) was connected to the output node. The learning rate was set to 0.001 and the momentum was set to 0.01. If these values are set too high, training fails to converge. The classification accuracy on the test set was observed to be 91.3% and on the training set it was 83%. Of the 221 resistant genes in the test set we were able to identify 202, while the remaining were mis-classified as susceptible genes.

4.1.2 Rule-Based Learning

The classifier generated by CN2 [6] is represented as a set of rules that describe the relationships between the plant attributes and the leaf miner resistance. CN2 induced 61 rules from the training set. Each rule "explains" a number of examples from the training set. This defines the *coverage* of the rule. CN2 tries to find rules that maximise the coverage for examples belonging to one class (the target class — resistant in this case) and minimise the coverage for all other classes (susceptible in this case). The first rule given below covers 114 examples belonging to the resistant class and 0 examples belonging to the susceptible class.

IF country = PAKISTHAN AND days to maturity < 188.50 AND seed weight < 30.65 THEN plant = "resistant" [114 0]

IF country = JORDAN AND biomass > 405.00 THEN plant = "resistant" [42 0]

The classification accuracy was 96.1% on the training set and 87.2% on the test set. Of the 221 resistant genes in the test set we were able to identify 206 genes. The remaining were mis-classified as susceptible genes.

5 Summary

Our study clearly demonstrates that we can identify/predict useful genes based on a carefully selected sub-set (core collection). However more research is required for selecting an appropriate machine learning system in crop genebanks. The research framework provided in this paper is currently underway at Queen's University.

References

[1] V.K.R. Addala. Evaluation of core collection strategies in chickpea (Cicer aritenum). Master's thesis, University of Saskatchewan, 1998.

[2] T.M Bataillon, J.L. David, and D.J Schoen. Neutral genetic markers and conservation genetics: Simulated germplasm collections. *Genetics*, 144:409-417, 1996.

[3] A.H.D. Brown. The case for core collections. In A.H.D. Brown, O.H. Frankel, D.R. Marshall, and J.T. Williams, editors, *The use of plant genetic resources*, pages 136-156. Cambridge University Press, Cambridge, UK, 1989.

[4] A.H.D. Brown. Core collections: A practical approach to genetic resources management. *Genome*, 31:818-824, 1989.

[5] A.H.D. Brown. The core collections at the crossroads. In T. Hodgkin, A.H.D. Brown, Th.J.L. van Hintum, and E.A.V. Morales, editors, *Core collections of plant genetic resource*, pages 3-19. John Wiley and Son, Chichester U.K., 1995.

[6] P. Clark and T. Niblett. The CN2 induction algorithm. *Machine Learning*, 3:261-283, 1988.

[7] National Research Council. *Managing global genetic resources: Agricultural crop issues and policies*. National Academy Press, Washington, D.C, US, 1993.

[8] N. Diwan, M.S. McIntosh, and G.R Bauchan. Methods of developing core collection of annual medicago species. *Theor. Appl. Genetics*, 90:755-761, 1995.

[9] W. Erskine and F.J. Muelbauer. Allozyme and morphological variability, outcrossing rate and core collection formation in lentil germplasm. *Theor. Appl. Genetics*, 83:119-125, 1991.

[10] FAO (Food and Agriculture Organization of the United Nations). *State of the world's plant genetic resources for food and agriculture*. FAO, Rome, Italy, 1996.

[11] O.H. Frankel. Genetic perspectives of germplasm collections. In W. Arber, K. Llimensee, W.J. Peacock, and P. Starlinger, editors, *Genetic manipulation: Impact on man and society*, pages 161-170. Cambridge University Press, Cambridge, UK, 1984.

[12] P.H. Goodman. Nevprop software, version 4. *NevProp3 User Manual*, University of Nevada, Reno, 1996.

[13] S. Hamon, M. Noirot, and F Anthony. Developing a coffee core collection using principal components score strategy with quantitative data. In T. Hodgkin, A.H.D. Brown, Th.J.L. van Hintum, and E.A.V. Morales, editors, *Core collections of plant genetic*

resource, pages 117-126. John Wiley and Son, Chichester U.K., 1995.

[14] J.T. Harlan. Genetic resources in sorghum. In *Sorghum in the seventies*, pages 1-13. Oxford and IBH, New Delhi, India, 1972.

[15] K.B. Singh, L. Holly, and G. Bejiga. *A catalog of kabuli chickpea germplasm.* ICARDA, Aleppo, Syria, 1991.

[16] J.P. Peeters and J.A Martinelli. Hierarchical cluster analysis as a tool to manage variation in germplasm collections. *Theor. Appl. Genetics*, 78:42-48, 1989.

[17] F. Provost, D. Jensen, and Tim Oates. Efficient progressive sampling. In *Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, CA, USA, August 15-18, 1999.

[18] D.J Schoen and A.H.D Brown. Maximizing genetic diversity in core collections of wild relatives of crop species. In T. Hodgkin, A.H.D. Brown, Th.J.L. van Hintum, and E.A.V. Morales, editors, *Core collections of plant genetic resource*, pages 55-76. John Wiley and Son, Chichester U.K., 1995.

[19] Th.J.L. van Hintum, R. van Bothmer, and D.L Visser. Sampling strategies for composing a core collection of cultivated barley (hordeum vulgare s. lat.). *Hereditas*, 122:7-17, 1995.

[20] D.A. Vaughan. Choosing rice germplasm for evaluation. *Euphytica*, 54:147-154, 1991.

[21] Spagnoletti Zeuli, P.L., and Qualset. Evaluation of five strategies for obtaining a core subset from a large genetic collection of durum wheat. *Theor. Appl. Genetics*, 87:295-304, 1993.

About the Authors

Raju Addala (http://www.cs.queensu.ca/home/addala) is a graduate student in the Department of Computing and Information Sciences, Queen's University, Kingston, ON K7L 3N6, Canada. His research interests include Machine Learning, Data mining, Sampling, Management & Utilization of genebanks, Bioinformatics and Population & Conservation genetics.

Ken Whelan is a Post-doctorate Fellow in the Computing and Information Science Department at Queen's University, Kingston Canada. His research interests are Machine Learning, Inductive Logic Programming and Knowledge Discovery in Molecular Biology. He received his Ph.D. in Machine Learning from the University of Abertay, Dundee, Scotland in 1998.

Janice Glasgow (http://www.qucis.queensu.ca/home/janice) is Professor and Head of the Department of Computing and Information Science, Queen's University, Kingston, K7L 3N6, Canada. Her research interests include computational imagery and its application to bioinformatics. Dr. Glasgow is a past President of the Canadian Society for Computational Studies of Intelligence, current NSERC Group Chair for Computing and Engineering Grant Selection Committees, a Director of the International Society of Computational Biology and Vice-Chair of the IFIP Technical Committee on Artificial Intelligence.

Some Transatlantic Observations About Graduate Research

Eli Hagen

Résumé

Je partagerai certaines de mes expériences en tant qu'étudiant gradué dans l'industrie, en Allemagne et comparerai ces dernières aux observations faites à deux universités canadiennes. Je considère certaines issues de surveillance, de ressources, et de collaboration.

Abstract

I will share some of my experiences as a "graduate student" in industry in Germany[1] and compare these with observations made at two Canadian Universities. I consider issues of supervision, resources, and collaboration.

Background

My academic background is in Computing Science (B.Sc. Queen's University, M.Sc. Simon Fraser University, and I am currently enrolled in the Ph.D. program at SFU). After my M.Sc., I left Canada to work at a research institute in Germany (GMD, German National Institute for Information Technology) and after two years there, I brought my research with me to Deutsche Telekom.

Research Infrastructure in Germany

The GMD is a national research institute. It is funded jointly by the government, the province (German: Bundesland), and industry. The industrial portion varies between 30 - 60% of their budget. The institutes are affiliated with a university and typically the head of the institute has a joint appointment as a professor[2].

At Deutsche Telekom, traditionally, 100% of research funding came from corporate funds. But funding is becoming increasingly more demand-driven, i.e., the research groups need to acquire a certain percentage of their budget from the product-related branches of Deutsche Telekom. Deutsche Telekom's research groups are not formally affiliated with a university, but many personal relations exist.

1. I will refer to this as the industry system. Other European countries have an educational system similar to the one described here, e.g. Norway.
2. There are several such institutes in Germany, e.g., DFKI, FAW, Forwiss, Frauenhofer Institute, and Max Planck Institute.

Graduate Research in Industry

While knowing that it was time for me to leave university after my M.Sc., I was not comfortable leaving research. Unfortunately, my search for a research related job in Canada was very discouraging, so I was excited when I was offered a job at the GMD. It was exactly what I wanted: I was doing research while gaining industry experience in a cooperative setting. I was not a student anymore, not working on the side to finance my studies, and there was no unguided search for a thesis topic. Certain areas became my areas of expertise and these I brought with me to a similar environment at Deutsche Telekom. The work I did for them became a solid basis from which I could start my Ph.D. thesis.

Observations

Our work was project-oriented and everybody worked toward a common goal, which encouraged cooperation. Also, a project manager with a deadline will pay much more attention to you than a supervisor with his/her own deadlines. The industry system seems to encourage more exchange of ideas and sometimes offer better supervision than a traditional professor-student relationship.

When going back to university I had one major advantage over non-industrial students: I did not have to spend several semesters aimlessly in search for a thesis topic. I can also capitalize on my previous experience and pull in results from my previous work when writing my thesis. As a result, I will be able to finish my degree within the expected time frame.

Another experience gained in industry is the opportunity to work with other organizations as a representative of one's own company. Such representational responsibilities are good for presenting and getting feedback on one's work, not to mention, a rich source of Vitamin C[3]. I represent Deutsche Telekom in an organization for standardization of my research area. The organization accepted me as a colleague and I am not so sure this would have been the case if I had gone there as a stu-

3. Based on a German expression, 'C' is for 'contacts'.

dent. And, who would have invited me to join if I were merely a student?

Another advantage of industry is that there are more funds available for equipment. For example, computer equipment, ergonomically correct desks and chairs, printing and photocopying facilities. I believe that good facilities increase a student's productivity.

Conclusion

In Germany, the industry system co-exists with the university system. The strength of their co-existence is that they reach more potential students. Some students know that they will pursue an academic career and for them, the university system works well. There are, however, many who need a break from the traditional system. For this group, the dual system offers an opportunity to take a job in the real world with the OPTION of turning their work into a thesis.

I have been lucky to be able to pursue a research career in an environment that fitted my needs. In order to achieve that, I was able and willing to move to a foreign country, but not everybody can do that. I wish it would be easier (or even possible) to do it in Canada — either in a dual system like in Germany or in a closer co-operation between universities and industry. I think graduate students could receive better and more relevant supervision in industry than from our already overworked professors. It would improve students' mental and financial health, and we may see shorter completion times and higher completion rates.

Since I do not suffer from thesis blues, am not broke, and did not spend 5+ years on my degree, I will be motivated, have several years of industry experience, some crucial vitamin C, and a degree when I will be looking for job — something, which I am not hearing from my fellow students.[4]

About the Author

Eli hagen is a Ph.D. candidate in Computing Science as Simon Fraser University, British Columbia. She received her B.Sc. in Computing Science from Queen's University, Ontario in 1991 and her M.Sc. in the same field from Simon Fraser University in 1993. Her primary research interests are spoken dialogue systems and text generation. Her industrial experience has been at GMD (German National Research Institute for Information Technology) and Deutsche Telekom. At both companies she was involved in computer human interaction research.

4. The reader may ask why I went back to the university, while singing the praises of industry. That was a strictly personal choice; I love the West Coast and going back to SFU for Ph.D. was a good opportunity to spend some mroe time here.

CONFERENCE REPORTS

RAPPORTS DE CONFÉRENCE

MT Summit in the Great Translation Era

Davide Turcato

Résumé

Le sommet au sujet de la traduction automatique est une conférence internationale discutant la TA sous tous ses angles. La 7ème édition a été tenue à Singapour en septembre dernier. Nous offrons un examen de la conférence, considérant la TA comme étant à la fois sujet de recherche et étude de cas des aspects politiques dans, l'espoir que ces deux aspects sont d'intérêt pour l'ensemble de la communauté de l'IA.

Abstract

The Machine Translation Summit is an international conference aimed at discussing MT from all perspectives. The 7th edition was held in Singapore last September. We offer a review of the conference, looking at MT as both a research topic and a case study of policy issues, in the belief that both aspects can be of interest to the AI community at large.

Conference Report

The Machine Translation Summit is a biennial conference that brings together MT researchers, developers, providers, users, policy makers and watchers to discuss all aspects of MT, from research advancements, to users' experiences, to policy issues. This year's edition, the seventh, was dedicated to the all-encompassing topic of "MT in the Great Translation Era" and took place from the 13th to the 17th of September at Kent Ridge Digital Labs, in warm and sunny Singapore (or S'pore, as the locals confidentially call it).

Far from being a sort of MT fair, the conference turned out an increasingly prominent event in the MT panorama, and an excellent opportunity for a stimulating discussion among all the different perspectives upon the field of MT.

Diverse Participants

The conference hosted a diversified audience, both geographically and in terms of professional backgrounds. The Table 1 gives an overview of the geographical distribution of the participants. The conference program was organized around 5 keynote speeches and invited talks, 8 special sessions, where invited papers were presented, 2 panels, 12 general sessions and 3 poster session, which were introduced for the first time in this edition. General and poster sessions comprised 67 contributed papers. Scheduled demos and permanent exhibitions completed the program.

	Delegates	Speakers	Countries
AFRICA	1	0	1
AMERICAS	30	7	2
ASIA	103	25	10
EUROPE	24	16	13
OCEANEA	8	1	1
	166	**49**	**27**

Table 1: Geographical Distribution of Participants based on the list of participants distributed at the conference. The organizers state that there were actually more than 250 participants. The 'Speakers' column refers to the invited speakers only.

Research Presentations

One of the most prominent strands of research displayed at the summit, and one which shows potential for cooperation with other sector of the AI research, is the application of MT to new domains and applications. A special session was dedicated to MT and speech, which included the presentation of a wearable computer usable by travellers exploring a foreign city, to provide translations and for the purpose of navigation (Waibel). Other applications included: the "translation camera", a camera system which translates Japanese texts in a scene, via a combination of character image extraction, character recognition and translation processes (Watanabe et al.); the "translating telephone" (Frederking et al.); a system to translate e-mail (Matsuda & Kumai), along with a number of systems for translating over the Web.

Besides addressing novel applications for the existing MT technology, another notable tendency is to address less ambitious fields related to MT, rather then MT in a strict sense. One such strand of research focuses on translation tools. A special session was dedicated to the topic. Several presentations dealt with translation memories and their integration with MT (e.g. Planas & Furuse and Carl & Hansen). Another tool that was presented is "TransRouter" (King), a support tool for translation managers, which helps making a decision about different translation options (human translation, use of specific terminology resources, translation memories, MT, etc.). Multilinguality is another topic, related to MT but distinct from it, that was widely discussed at the Summit. Two general sessions were dedicated to "multilingual human language technology" and "multilingual information processing", and a special session was dedicated to "multilingual information access"; among other applications, multilingual information retrieval and multilingual language recognition were discussed.

To conclude this incomplete overview, we would like to mention a few topics more directly related to the task of MT, as traditionally intended. One such topic was the automatic or semi-automatic development of lexical resources. A session was dedicated to "knowledge extraction" and several other papers dealt with the topic. In terms of MT methodologies, a session was devoted to Example-based MT and another one to "statistical methods".

Growing Demand for Low Quality?

In general, the impression that one reports is that most of the research effort goes in the direction of making the best of the existing resources and technology, rather than investing in the long term goal of high-quality translation. This situation has been summarized by Martin Kay in his invited talk, in the following terms:

"If we somehow seem to be winning the game that has been going against us for so long, it is not because we have learned to play better; it is because the goal posts have been moved. Simply stated, the market for low-quality machine translation has grown from nothing to one clearly worthy of commercial interest in a matter of two or three years."

Martin Kay's presentation was indeed the most notable exception to the tendency towards trying to make the best of the current level of translation quality. He made a stimulating presentation of an approach to translation called "chart translation", based on systematically preserving ambiguities.

Policy Matters

This hint to the contrast between the long-term goal of high-quality translation and the requirement that investments be profitable brings us to another prominent issue that was discussed at the MT Summit, i.e. policy making. MT is probably one of the most instructive case studies, in this respect, for the entire AI community. On the one hand, MT has an enormous potential for commercial exploitation and has always been looked upon with high expectations. On the other hand, it is known to be an "AI-complete" problem, which is far from being satisfactorily solved.

At the Summit there was an interesting panel about "International Cooperation: From Funding Agencies' Point of View", in which, among others, representatives of the European Commission and DARPA took part. Both representatives ended up admitting that investments in MT are done with great caution, since no major breakthrough in the field has come to justify a greater optimism. On the other hand, an indirect response from the research point of view came from an interesting presentation by Christian Boitet, about "how to democratize machine translation and translation aids aiming at high quality final output". Boitet argued that a major breakthrough, within the current technological horizon, could precisely come from different policies, aiming at a wider sharing of existing resources, as a way of addressing the aforementioned AI-completeness of MT. Another interesting outlook on policy issues, from the user's perspective, was provided by Dimitri Theologitis, of the EU Translation Services Bureau, who stated that, to some extent, human translators are used by the EU for policy reasons, whereas in some cases the use of MT would be more cost-effective.

Last but not least, we would like to acknowledge that the food at the MT Summit was good and abundant, very much in the Singaporean tradition, even for coffee breaks. Someone was heard draw comparisons with last year's COLING-ACL, but it is not known in what terms.

Machine Translation 2001

The next edition of the MT Summit will be held from the 18th to the 22nd September 2001 in Spain, in the spiritual and administrative capital of Galicia, Santiago de Compostela, according to a scheme that alternates Asia, Europe and (so far) United States. The Galician local organizers were in Singapore, and we checked with them the most pressing issues. They reassured us: the food is great there too (excellent seafood and fish), and they also have good wines. More information about the MT Summit can be found at http://www.jeida.or.jp/aamt/mtsummit99/index.html.

About the Author

Davide Turcato (turk@cs.sfu.ca) is a research associate in the Natural Language Lab, School of Computer Science, Simon Fraser University, Burnaby Mountain, BC, V5A 1S6, Canada. His research interests include, from more general to more specific, the formalization of natural languages, Natural Language Processing, Machine Translation, Computer Aided Language Learning, Information Retrieval.

The Tenth Annual PRECARN • IRIS Conference

May 14-17, 2000

Palais des Congrès, Montréal, QC

Mark your calendars!

We're very excited about this year's conference. Not only is it the Tenth PRECARN • IRIS Annual Conference but the Conference is also being held for the first-time in conjunction with the 31st International Symposium on Robotics (ISR 2000), and the Artificial Intelligence (AI) 2000, Graphics Interface (GI) 2000 and Vision Interface (VI) 2000 conferences. Five Conferences for the price of one!! We are hoping to attract over 800 people from around the world.

You won't want to miss Canadian Astronaut, Julie Payette, who will launch this first-time event at the Opening Plenary session. An accomplished computer science researcher, Ms. Payette will be sharing with us reports of her voyage into space as part of the first shuttle mission to dock with the International Space Station (ISS).

As a favorite of the Conference in the past years the Demos and Posters exhibitions return. These exhibits will provide a unique opportunity for conference participants to see, touch and learn about PRECARN and IRIS ongoing research projects.

For conference information, contact:

Conference Coordinator Office
Golden Planners Inc.
126 York Street, Suite 301
Ottawa, Ontario K1N 5T5
TEL (613) 241-9333 / FAX (613) 565-2173
EMAIL ISR2000@goldenplanners.com

For conference information updates please visit our web site	
ISR 2000	**PRECARN•IRIS Conference**
www.precarn.ca/isr2000	**www.precarn.ca/Events**

CSCSI/SCEIO Membership Application

I wish to join CSCSI/SCEIO and receive *Canadian Artificial Intelligence /Intelligence Artificielle au Canada*

Yes ___ / No ___ Web Access Only ($30.00* Cdn./yr.)

Yes ___ / No ___ Printed Copy and Web Access ($40.00* Cdn./yr.)

I am a student

Yes ___ / No ___ Web Access Only ($15.00* Cdn./yr.)

Yes ___ / No ___ Printed Copy and Web Access ($15.00* Cdn./yr.)

I am a member of CIPS

Yes ___ / No ___ Web Access Only ($25.00* Cdn./yr.)

Yes ___ / No ___ Printed Copy and Web Access ($30.00* Cdn./yr.)

*Includes applicable G.S.T.

Name ....

Job Title ....

Company ....

Address ....

....

....

....

E-mail ....

Phone ....

Fax ....

Visa or Amex Number ....

Expiry Date ....

Signature ....

Please send your membership application to:

Mail: CIPS National Office, One Yonge St., Suite 2401, Toronto, Ontario M5E 1E5

Phone: (416) 861-2477

Fax: (416) 368-9972

E-mail: aff@cips.ca

For more information, contact CIPS or a member of the CSCSI/SCEIO executive.

Canadian Artificial Intelligence
Intelligence Artificielle au Canada

The editor welcomes contributions from members and non-members. Possible formats include, but are not limited to:

- commentary, in the form of letters or short articles, on any AI issue of concern to you (e.g. research priorities, funding, educational trends, cultural, political, or ethical aspects of AI)
- interdisciplinary work
- survey papers
- reports on new work by students
- abstracts
- web site reviews
- news
- humour
- employment opportunities
- conference announcements.

Submission information • Guide de soumission

Canadian Artificial Intelligence / Intelligence Artificielle au Canada welcomes submissions on any matter related to artificial intelligence. Please send your contribution, preferably in electronic form, with an abstract and biography to:

Ann Grbavec
Editor, *Canadian AI / IA au Canada*
School of Computing Science
Simon Fraser University
Burnaby, British Columbia, Canada
V5A 1S6 or — agrbavec@cs.sfu.ca
Phone (604) 291-3208
Fax (604) 291-3045

Please contact the editor as early as possible to obtain complete submission guidelines.
Submission deadline: February 28, 2000.

Sample issues and articles can be viewed at: http://cscsi.sfu.ca/cai.html

Advertising rates are available upon request from the editor.

Recycled / Recyclable