TiMBL: Tilburg Memory Based Learner
version 1.0
Reference Guide

ILK Technical Report – ILK 98-03

Walter Daelemans Jakub Zavrel Ko van der Sloot
Antal van den Bosch

Induction of Linguistic Knowledge
Computational Linguistics
Tilburg University
P.O. Box 90153, NL-5000 LE, Tilburg, The Netherlands
URL: http://ilk.kub.nl[1]

March 17, 1998

[1]This document is available from http://ilk.kub.nl/~ilk/papers/ilk9803.ps.gz.

Contents

Preface

Memory-Based Learning (MBL) has proven to be quite successful in a large number of tasks in Natural Language Processing (NLP). In our group at Tilburg University we have been working since the end of the 1980's on the development of Memory-Based Learning techniques and algorithms[1]. With the establishment of the ILK (Induction of Linguistic Knowledge) research group in 1997, the need for a well-coded and uniform tool for our main algorithms became more urgent. TiMBL is the result of combining ideas from a number of different MBL implementations, cleaning up the interface, and using a whole bag of tricks to make it more efficient. We think it can make a useful tool for NLP research, and, for that matter, for all other domains with discrete classification tasks.

Memory-Based Learning is a direct descendant of the classical k-Nearest Neighbor (k-NN) approach to classification. In typical NLP learning tasks, however, the focus is on discrete data, very large numbers of examples, and many attributes of differing relevance. Moreover, classification speed is a critical issue in any realistic application of Memory-Based Learning. These constraints, which are quite different from those of traditional pattern recognition applications with their numerical features, often lead to different data-structures and different speedup optimizations for the algorithms. Our approach has resulted in an architecture which makes extensive use of indexes into the instance memory, rather than the typical flat file organization found in straightforward k-NN implementations. In some cases the internal organization of the memory results in algorithms which are quite different from k-NN, as is the case with IGTREE. We believe that our optimizations make TiMBL one of the fastest discrete k-NN implementations around.

The main effort in the development of this software was done by Ko van der Sloot. The code started as a rewrite of `nibl`, a piece of software developed by Peter Berck from a Common Lisp implementation by Walter Daelemans. Some of the index-optimizations are due to Jakub Zavrel. The code has benefited substantially from trial, error and scrutiny by the other members of the ILK group (Sabine Buchholz, Jorn Veenstra and Bertjan Busser). We would also like to thank Ton Weijters of the Technical University of Eindhoven and the members

[1] Section 3.5 provides a historical overview of our work on the application of MBL in NLP.

of the CNTS research group at the University of Antwerp for their contributions. This software was written in the context of the "Induction of Linguistic Knowledge" research programme, partially supported by the Foundation for Language Speech and Logic (TSL), funded by the Netherlands Organization for Scientific Research (NWO).

The current release (version 1.0) is a first beta release and although it was tested for some time in our research group, it may still contain bugs and inconsistencies in certain places. This reference guide is also a first version. We would appreciate it if you can send bug reports, ideas about enhancements of the software and the manual, and any other comments you might have, to `Timbl@kub.nl`.

This reference guide is structured as follows. In Chapter 1 you can find the terms of the license according to which you are allowed to use TiMBL. The following chapter gives some instructions on how to install the TiMBL package on your computer. Readers who are interested in the theoretical and technical details of Memory-Based Learning and of this implementation can then proceed to Chapter 3. Those who just want to get started using TiMBL can skip this chapter, and directly proceed either to Chapters 4 and 5, which respectively provide a reference to the file formats and command line options of TiMBL, or to Appendix A, where a short hands-on tutorial is provided on the basis of a case study with a data set from a linguistic domain (prediction of Dutch diminutive suffixes).

Chapter 1

License terms

Downloading and using the TiMBL software implies that you accept the following license terms:

Tilburg University grants you (the registered user) the non-exclusive license to download a single copy of the TiMBL program code and related documentation (henceforth jointly referred to as "Software") and to use the copy of the code and documentation solely in accordance with the following terms and conditions:

- The license is only valid when you register as a user. If you have obtained a copy without registration, you must immediately register by sending an e-mail to `Timbl@kub.nl`.
- You may only use the Software for educational or non-commercial research purposes.
- You may make and use copies of the Software internally for your own use.
- Without executing an applicable commercial license with Tilburg University, no part of the code may be sold, offered for sale, or made accessible on a computer network external to your own or your organization's in any format; nor may commercial services utilizing the code be sold or offered for sale. No other licenses are granted or implied.
- Tilburg University has no obligation to support the Software it is providing under this license. To the extent permitted under the applicable law, Tilburg University is licensing the Software "AS IS", with no express or implied warranties of any kind, including, but not limited to, any implied warranties of merchantability or fitness for any particular purpose or warranties against infringement of any proprietary rights of a third party and will not be liable to you for any consequential, incidental, or special damages or for any claim by any third party.

- Under this license, the copyright for the Software remains the property of the ILK Research Group at Tilburg University. Except as specifically authorized by the above licensing agreement, you may not use, copy or transfer this code, in any form, in whole or in part.
- Tilburg University may at any time assign or transfer all or part of its interests in any rights to the Software, and to this license, to an affiliated or unaffiliated company or person.
- Tilburg University shall have the right to terminate this license at any time by written notice. Licensee shall be liable for any infringement or damages resulting from Licensee's failure to abide by the terms of this License.
- In publication of research that makes use of the Software, a citation should be given of: *"Walter Daelemans, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch (1998). TiMBL: Tilburg Memory Based Learner, version 1.0, Reference Guide. ILK Technical Report 98-03, Available from* `http://ilk.kub.nl/~ilk/papers/ilk9803.ps.gz`
- For information about commercial licenses for the Software, contact `Timbl@kub.nl`, or send your request in writing to:

Dr. Walter Daelemans
ILK Research Group
Computational Linguistics
Tilburg University
PO Box 90153
5000 LE Tilburg
The Netherlands

Chapter 2

Installation

You can get the TiMBL package as a gzipped tar archive from:

`http://ilk.kub.nl/software.html`

Following the links from that page, you will be required to fill in registration information and to accept the license agreement. You can then proceed to download the file `Timbl.1.0.tar.gz`
This file contains the complete source code (C++) for the TiMBL program, a few sample data sets, the license and the documentation. The installation should be relatively straightforward on most UNIX systems.

To install the package on your computer, unzip the downloaded file:

```
> gunzip Timbl.1.0.tar.gz
```

and unpack the tar archive:

```
> tar -xvf Timbl.1.0.tar
```

This will make a directory `Timbl.1.0` under your current directory. Change directory to this:

```
> cd Timbl.1.0
```

and compile the executable by typing `make`[1]. If the process was completed successfully, you should now have an executable file named `Timbl`.

The e-mail address for problems with the installation, bug reports, comments and questions is `Timbl@kub.nl`.

[1] We have tested this only with `gcc` version 2.7.2

Chapter 3

Learning algorithms

TiMBL is a program implementing several Memory-Based Learning techniques. All the algorithms have in common that they store some representation of the training set explicitly in memory. During testing, new cases are classified by extrapolation from the most similar stored cases. The main differences between the algorithms incorporated in TiMBL lie in:

- The definition of *similarity*,
- The way the instances are stored in memory, and
- The way the search through memory is conducted.

In this chapter, various choices for these issues are described. We start in section 3.1 with a formal description of the basic Memory-Based Learning algorithm, i.e. a nearest neighbor search. We then introduce different similarity metrics, such as Information Gain weighting, which allows us to deal with features of differing importance, and the Modified Value Difference metric, which allows us to make a graded guess of the match between two different symbolic values. In section 3.2 and 3.3, we give a description of various optimizations for nearest neighbor search. Finally, in section 3.4, we describe the fastest optimization, IGTREE, which replaces the exact nearest neighbor search with a very fast heuristic that exploits the difference in importance between features.

3.1 Memory Based Learning

Memory-based learning is founded on the hypothesis that performance in cognitive tasks is based on reasoning on the basis of similarity of new situations to *stored representations of earlier experiences*, rather than on the application of *mental rules* abstracted from earlier experiences (as in rule induction and rule-based processing). The approach has surfaced in different contexts using a variety of alternative names such as similarity-based, example-based, exemplar-based, analogical, case-based, instance-based, and lazy learning [22, 5, 19, 2, 1].

Historically, memory-based learning algorithms are descendants of the k-nearest neighbor (henceforth k-NN) algorithm [6, 16, 2].

An MBL system, visualized schematically in Figure 3.1, contains two components: a *learning component* which is memory-based (from which MBL borrows its name), and a *performance component* which is similarity-based.

The learning component of MBL is memory-based as it involves adding training instances to memory (the *instance base* or case base); it is sometimes referred to as 'lazy' as memory storage is done without abstraction or restructuring. An instance consists of a fixed-length vector of n feature-value pairs, and an information field containing the classification of that particular feature-value vector.

In the performance component of an MBL system, the product of the learning component is used as a basis for mapping input to output; this usually takes the form of performing classification. During classification, a previously unseen test example is presented to the system. The similarity between the new instance X and all examples Y in memory is computed using a *distance metric* $\Delta(X, Y)$. The extrapolation is done by assigning the most frequent category within the k most similar example(s) as the category of the new test example.

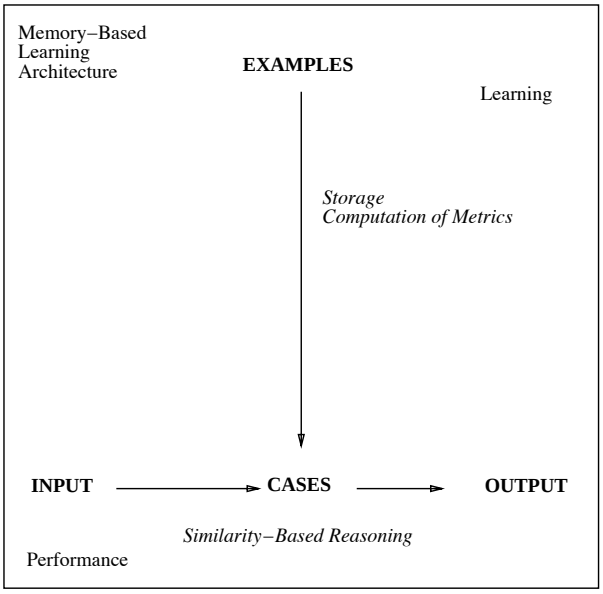

Figure 3.1: General architecture of an MBL system.

3.1.1 Overlap metric

The most basic metric for patterns with symbolic features is the **Overlap metric** given in equations 3.1 and 3.2; where $\Delta(X, Y)$ is the distance between

patterns X and Y, represented by n features, and δ is the distance per feature. The distance between two patterns is simply the sum of the differences between the features. The k-NN algorithm with this metric is called IB1 [2]. Usually k is set to 1.

$$\Delta(X, Y) = \sum_{i=1}^{n} \delta(x_i, y_i) \tag{3.1}$$

where:

$$\delta(x_i, y_i) = \begin{cases} 0 & \text{if } x_i = y_i \\ 1 & \text{if } x_i \neq y_i \end{cases} \tag{3.2}$$

We have made two additions to the original algorithm [2] in our version of IB1. First, in the case of nearest neighbor sets larger than one instance ($k > 1$ or ties), our version of IB1 selects the classification that has the highest frequency in the class distribution of the nearest neighbor set. Second, if a tie cannot be resolved in this way because of equal frequency of classes among the nearest neighbors, the classification is selected with the highest overall occurrence in the training set.

3.1.2 Information Gain weighting

The distance metric in equation 3.2 simply counts the number of (mis)matching feature-values in both patterns. In the absence of information about feature relevance, this is a reasonable choice. Otherwise, we can add domain knowledge bias to weight or select different features (see e.g. Cardie [4] for an application of linguistic bias in a language processing task), or look at the behavior of features in the set of examples used for training. We can compute statistics about the relevance of features by looking at which features are good predictors of the class labels. Information Theory gives us a useful tool for measuring feature relevance in this way [20, 21].

Information Gain (IG) weighting looks at each feature in isolation, and measures how much information it contributes to our knowledge of the correct class label. The Information Gain of feature i is measured by computing the difference in uncertainty (i.e. entropy) between the situations without and with knowledge of the value of that feature (equation 3.3).

$$w_i = H(C) - \sum_{v \in V_i} P(v) \times H(C|v) \tag{3.3}$$

Where C is the set of class labels, V_i is the set of values for feature i, and $H(C) = -\sum_{c \in C} P(c) \log_2 P(c)$ is the entropy of the class labels. The probabilities are estimated from relative frequencies in the training set.

It is important to realize that the IG weight is really a probability weighted average of the informativity of the different values of the feature. On the one hand, this pre-empts the consideration of values with low frequency but high informativity. Such values "disappear" in the average. On the other hand, this

also makes the IG weight very robust to estimation problems. Each parameter (=weight) is estimated on the whole data set.

Information Gain, however, tends to overestimate the relevance of features with large numbers of values. Imagine a data set of hospital patients, where one of the available features is a unique "patient ID number". This feature will have very high Information Gain, but it does not give any generalization to new instances. To normalize Information Gain for features with different numbers of values, Quinlan [21] has introduced a normalized version, called Gain Ratio, which is Information Gain divided by $si(i)$ (split info), the entropy of the feature-values (equation 3.5).

$$w_i = \frac{H(C) - \sum_{v \in V_i} P(v) \times H(C|v)}{si(i)} \tag{3.4}$$

$$si(i) = -\sum_{v \in V_i} P(v) \log_2 P(v) \tag{3.5}$$

The resulting Gain Ratio values can then be used as weights w_f in the weighted distance metric (equation 3.6)[1]. The k-NN algorithm with this metric is called IB1-IG [8].

$$\Delta(X, Y) = \sum_{i=1}^{n} w_i \, \delta(x_i, y_i) \tag{3.6}$$

The possibility of automatically determining the relevance of features implies that many different and possibly irrelevant features can be added to the feature set. This is a very convenient methodology if domain knowledge does not constrain the choice enough beforehand, or if we wish to measure the importance of various information sources experimentally. However, because IG values are computed for each feature independently, this is not necessarily the best strategy. Sometimes better results can be obtained by leaving features out than by letting them in with a low weight. Very redundant features can also be challenging for IB1-IG, because IG will overestimate their joint relevance. Imagine an informative feature which is duplicated. This results in an overestimation of IG weight by a factor two, and can lead to accuracy loss, because the doubled feature will dominate the similarity metric.

3.1.3 Modified Value Difference metric

It should be stressed that the choice of representation for instances in MBL remains the key factor determining the strength of the approach. The features and categories in NLP tasks are usually represented by symbolic labels. The metrics that have been described so far, i.e. Overlap and IG Overlap, are limited to exact match between feature-values. This means that all values of a feature

[1] In a generic use IG refers both to Information Gain and to Gain Ratio throughout this manual. In specifying parameters for the software, the distinction between both needs to be made, because they often result in different behavior.

are seen as equally dissimilar. However, if we think of an imaginary task in e.g. the phonetic domain, we might want to use the information that 'b' and 'p' are more similar than 'b' and 'a'. For this purpose a metric was defined by Stanfill & Waltz [22] and further refined by Cost & Salzberg [5]. It is called the (Modified) Value Difference Metric (MVDM; equation 3.7), and it is a method to determine the similarity of the values of a feature by looking at co-occurrence of values with target classes. For the distance between two values V_1, V_2 of a feature, we compute the difference of the conditional distribution of the classes C_i for these values.

$$\delta(V_1, V_2) = \sum_{i=1}^{n} |P(C_i|V_1) - P(C_i|V_2)| \tag{3.7}$$

For computational efficiency, all pairwise $\delta(V_1, V_2)$ values can be computed before the actual nearest neighbor search starts.

Although the MVDM metric does not explicitly compute feature relevance, an implicit feature weighting effect is present. If features are very informative, their conditional class probabilities will on average be very skewed towards a particular class. This implies that on average the $\delta(V_1, V_2)$ will be large. For uninformative features, on the other hand, the conditional class probabilities will be pretty uniform, so that on average the $\delta(V_1, V_2)$ will be very small.

MVDM differs considerably from Overlap based metrics in its composition of the nearest neighbor sets. Overlap causes an abundance of ties in nearest neighbor position. For example, if the nearest neighbor is at a distance of one mismatch from the test instance, then the nearest neighbor set will contain the entire partition of the training set that matches all the other features but contains *any* value for the mismatching feature (see [27] for a more detailed discussion). With the MVDM metric, however, the nearest neighbor set will only contain patterns which have the value with the lowest $\delta(V_1, V_2)$ in the mismatching position[2]. In sum, this means that the nearest neighbor set is usually much smaller for MVDM at the same value of k. In NLP tasks we have found it very useful to experiment with values of k larger than one for MVDM, because this re-introduces some of the beneficial smoothing effects associated with large nearest neighbor sets.

One cautionary note about this metric is connected to data sparsity. In many practical applications, we are confronted with a very limited set of examples. This poses a serious problem for the MVDM metric. Many values occur only once in the whole data set. This means that if two such values occur with the same class the MVDM will regard them as identical, and if they occur with two different classes their distance will be maximal. The latter condition reduces the MVDM to the Overlap metric for many cases, with the addition that some cases will be counted as an exact match or mismatch on the basis of very shaky evidence.

[2]Or MVDM will select a totally different nearest neighbor which has less exactly matching features, but a smaller distance in the mismatching feature.

3.2 Tree-based memory

The discussion of the algorithm and the metrics in the section above is based on a naive implementation of nearest neighbor search: a flat array of instances which is searched from beginning to end while computing the similarity of the test instance with each training instance (see the left part of Figure 3.2). Such an implementation, unfortunately, reveals the flip side of the lazy learning coin. Although learning is very cheap: just storing the instances in memory, the computational price of classification can become very high for large data sets. The computational cost is proportional to N, the number of instances in the training set.

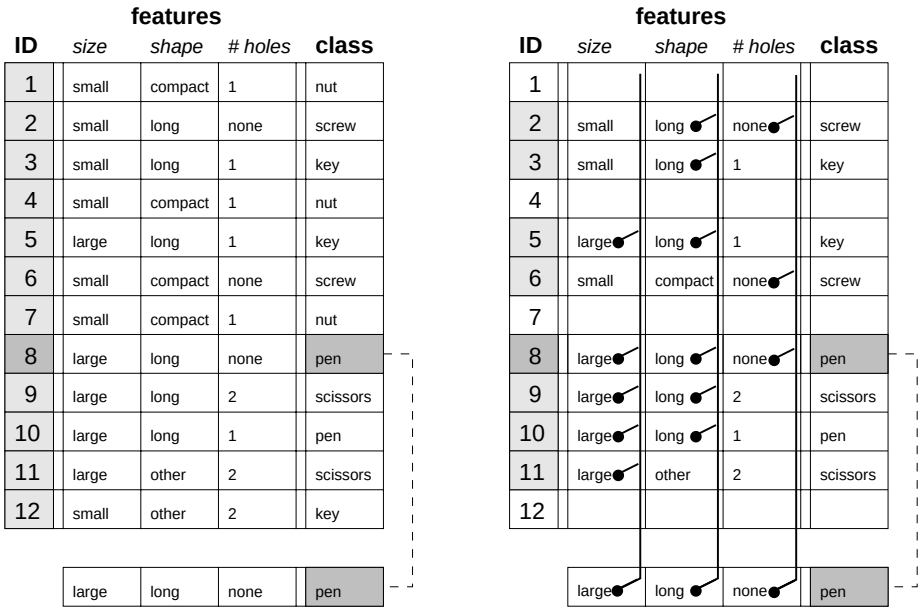

Figure 3.2: The instance base for a small object classification toy problem. The left figure shows a flat array of instances through which sequential nearest neighbor search is performed to find the best match for a test instance (shown below the instance base). In the right part, an inverted index (see text) is used to restrict the search to those instances which share at least one feature value with the test instance.

In our implementation of MBL we use a more efficient approach. The first part of this approach is to replace the flat array by a tree-based data structure. Instances are stored in the tree as paths from a root node to a leaf, the arcs of the path are the consecutive feature-values, and the leaf node contains a *distribution* of classes, i.e. a count of how many times which class occurs with this pattern of feature-values (see Figure 3.3).

Due to this storage structure, instances with identical feature-values are

collapsed into one path, and only their separate class information needs to be stored in the distribution at the leaf node. Many different **tokens** of a particular **instance type** share one path from the root to a leaf node. Moreover, instances which share a prefix of feature-values, also share a partial path. This reduces storage space (although at the cost of some book-keeping overhead) and has two implications for nearest neighbor search efficiency.

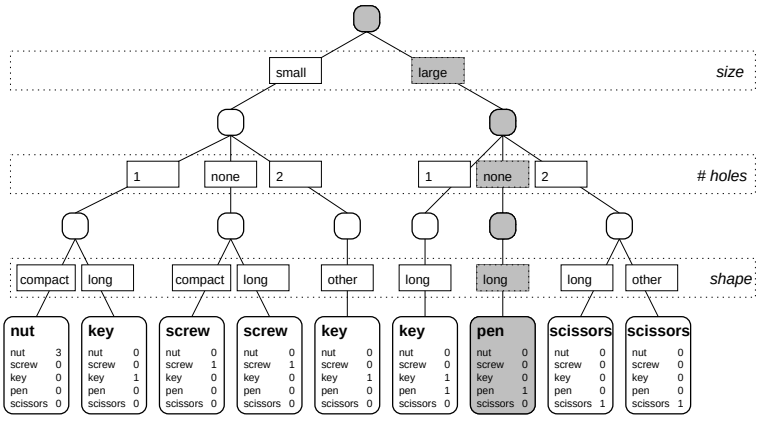

Figure 3.3: A tree-structured storage of the instance base from figure 3.2. An exact match for the test is in this case directly found by a top down traversal of the tree (grey path). If there is no exact match, all paths are interpreted as instances and the distances are computed. The order of the features in this tree is based on Gain Ratio.

In the first place, the tree can be searched top-down very quickly for *exact matches*. Since an exact match ($\Delta(X, Y) = 0$) can never be beaten, we choose to omit any further distance computations when one is found with this shortcut[3].

Second, the distance computation for the nearest neighbor search can re-use partial results for paths which share prefixes. This re-use of partial results is in the direction from the root to the leaves of the tree. When we have proceeded to a certain level of the tree, we know how much similarity (equation 3.2) can still contribute to the overall distance (equation 3.1), and discard whole branches of the tree which will never be able to rise above the partial similarity of the current least similar best neighbor.

Disregarding this last restriction, the number of feature-value comparisons is equal to the number of arcs in the tree. Thus if we can find an ordering of the features which produces more overlap between partial paths, and hence a smaller tree, we can gain both space and time improvements. An ordering which was found to produce small trees for many of our NLP data sets is Gain Ratio divided by the number of feature-values (this is the default setting). Through

[3]There is a command line switch (`-x`) which turns the shortcut off in order to get exact results when $k > 1$ (i.e. get neighbors at further distances).

the -T command line switch, however, the user is allowed to experiment with different orderings.

3.3 Inverse index

The second part of our approach to efficiency is a speedup optimization based on the following fact. Even in the tree-based structure, the distance is computed between the test instance and *all* instance types. This means that even instance types which do not share a single feature-value with the test instance are considered, although they will surely yield a zero similarity. The use of an **inverted index** excludes these zero similarity patterns. The construction of the inverted index records for all values of each feature a list of instance types (i.e. leaf nodes in the tree described in the previous section) in which they occur. Thus it is an inverse of the instance-base, which records for each instance type which feature-values occur in it[4].

When a test instance is to be classified, we select the lists of instance types for the feature-values that it contains (illustrated in the rightmost part of Figure 3.2). We can now find the nearest neighbor in these lists in a time that is proportional to the number of occurrences of the most frequent feature-value of the test pattern, instead of proportional to the number of instance types.

Although worst case complexity is still proportional to N, the size of the training set, and practical mileage may vary widely depending on the peculiarities of your data, the combination of exact match shortcut, tree-based path re-use, and inverted index has proven in practice (for our NLP datasets) to make the difference between hours and seconds of computation[5].

3.4 IGTree

Using Information Gain rather than unweighted Overlap distance to define similarity in IB1 improves its performance on several NLP tasks [8, 24, 23]. The positive effect of Information Gain on performance prompted us to develop an alternative approach in which the instance memory is restructured in such a way that it contains the same information as before, but in a compressed decision tree structure. We call this algorithm IGTREE [13] (see Figure 3.4 for an illustration). In this structure, similar to the tree-structured instance base described above, instances are stored as paths of connected nodes which contain classification information. Nodes are connected via arcs denoting feature values.

[4] Unfortunately this also implies that the storage of both an instance-base and an inverted index takes about twice the amount of memory.

[5] Due to the reasons described in footnote 2, MVDM cannot make use of the inverted index optimization. Because the precomputation of differences between values is often impossible in tasks with a large number of feature values (n^2 differences must be stored per feature, if n is the number of values of that feature), and because MVDM then effectively multiplies the number of distance computations per instance by the number of classes, this metric is currently one of the slowest in the package.

Information Gain is used to determine the order in which instance feature-values are added as arcs to the tree. The reasoning behind this compression is that when the computation of information gain points to one feature clearly being the most important in classification, search can be restricted to matching a test instance to those memory instances that have the same feature-value as the test instance at that feature. Instead of indexing all memory instances only once on this feature, the instance memory can then be optimized further by examining the second most important feature, followed by the third most important feature, etc. Again, considerable compression is obtained as similar instances share partial paths.

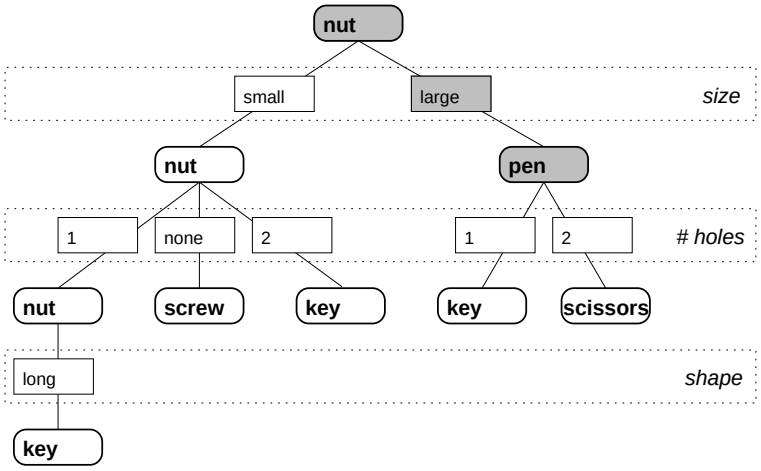

Figure 3.4: A pruned IGTREE for the instance base of Figure 3.2. The classification for the test instance is found by top down search of the tree, and returning the class label (default) of the node after the last matching feature-value (arc). Note that this tree is essentially a compressed version of the tree in Figure 3.3.

Because IGTREE makes a heuristic approximation of nearest neighbor search by a top down traversal of the tree in the order of feature relevance, we no longer need to store all the paths. The idea is that it is not necessary to fully store those feature-values of the instance that have lower Information Gain than those features which already fully disambiguate the instance classification.

Apart from compressing all training instances in the tree structure, the IGTREE algorithm also stores with each non-terminal node information concerning the *most probable* or *default* classification given the path thus far, according to the bookkeeping information maintained by the tree construction algorithm. This extra information is essential when processing unknown test instances. Processing an unknown input involves traversing the tree (i.e., matching all feature-values of the test instance with arcs in the order of the overall feature Information Gain), and either retrieving a classification when a leaf is reached

(i.e., an exact match was found), or using the default classification on the last matching non-terminal node if an exact match fails.

In sum, it can be said that in the trade-off between computation during learning and computation during classification, the IGTREE approach chooses to invest more time in organizing the instance base using Information Gain and compression, to obtain considerably simplified and faster processing during classification, as compared to IB1 and IB1-IG.

The generalization accuracy of IGTREE is usually comparable to that of IB1-IG; most of the time not significantly differing, and occasionally slightly (but statistically significantly) worse, or even better. The two reasons for this surprisingly good accuracy are that (i) most 'unseen' instances contain considerably large parts that fully match stored parts of training instances, and (ii) the probabilistic information stored at non-terminal nodes (i.e., the default classifications) still produces strong 'best guesses' when exact matching fails. The difference between the top-down traversal of the tree and precise nearest neighbor search becomes more pronounced when the differences in informativity between features are small. In such a case a slightly different weighting would have produced a switch in the ordering and a completely different tree. The result can be a considerable change in classification outcomes, and hence also in accuracy. However, we have found in our work on NLP datasets that when the goal is to obtain a very fast classifier for processing large amounts of text, the slight tradeoff between accuracy and speed can be very attractive.

3.5 NLP applications of TiMBL

This section provides a historical overview of our own work with the application of MBL type algorithms to NPL tasks.

The IB1-IG algorithm was first introduced in [8] in the context of a comparison of memory-based approaches with backprop learning for a hyphenation task. Predecessor versions of IGTREE can be found in [10, 24] where they are applied to grapheme-to-phoneme conversion. See [13] for a detailed description and review of the algorithms. A recent development, not yet implemented in the TiMBL package is TRIBL [14], an algorithm which constitutes a hybrid between the IB1-IG and IGTREE algorithms.

The memory-based algorithms implemented in the TiMBL package have been successfully applied to a large range of Natural Language Processing tasks: hyphenation and syllabification ([8]); assignment of word stress ([9]); grapheme-to-phoneme conversion ([11]); diminutive formation ([15]); morphological analysis ([25]); part of speech tagging ([12]); PP-attachment ([28]). Not yet published experimental results exist for word sense disambiguation, subcategorisation, and chunking (partial parsing).

Relations to statistical language processing are discussed in [27]. A partial overview paper is [7]. The first dissertation-length study devoted to the approach is [23], in which the approach is compared to alternative learning methods for NLP tasks related to English word pronunciation (stress assign-

ment, syllabification, morphological analysis, alignment, grapheme-to-phoneme conversion).

All papers referred to in this section are available in electronic form from the ILK homepage: `http://ilk.kub.nl`. We are grateful for any feedback on the algorithms and the way we applied them.

Whereas the work in Tilburg has been oriented primarily towards *language engineering* applications, the CNTS research group of Antwerp University, with which close research ties exist, has studied the linguistic and psycholinguistic relevance of memory-based learning for stress assignment in Dutch ([9, 18]), and as a model for *phonological bootstrapping*. A recently started project has as its aim to test predictions from memory-based models for language processing with psycholinguistic experiments.

Chapter 4

File formats

This chapter describes the format of the input and output files used by TiMBL. Where possible, the format is illustrated using the same small toy data set that is shown in Figure 3.2. It consists of 12 instances of 5 different everyday objects (nut, screw, key, pen, scissors), described by 3 discrete features (size, shape, and number of holes).

4.1 Data format

The training and test sets for the learner consist of descriptions of instances in terms of a fixed number of feature-values. TiMBL supports a number of different formats for the instances, but they all have in common that the files should contain one instance per line. The number of instances is determined automatically, and the format of each instance is inferred from the format of the first line in the training set. The last feature of the instance is assumed to be the target category. Should the guess of the format by TiMBL turn out to be wrong, you can force it to interpret the data as a particular format by using the `-F` option. Note that TiMBL is designed to deal with *symbolic, discrete values*, and that it will **not** interpret numbers as such but as just another string of characters.

Once TiMBL has determined the input format, it will skip and complain about all lines in the input which do not respect this format (i.e. have a different number of feature-values with respect to that format).

During testing, TiMBL writes the classifications of the test set to an output file. In most cases, the format of this output file is the same as the input format, with the addition of the predicted category being appended after the correct category.

4.1.1 Column format

The **column format** uses white space as the separator between features. White space is defined as a sequence of one or more spaces or tab characters. Every instance of white space is interpreted as a feature separator, so it is not possible to have feature-values containing white space. The column format is auto-detected when an instance of white space is detected on the first line *before a comma has been encountered*. The example data set looks like this in the column format:

```
small compact 1 nut
small long none screw
small long 1 key
small compact 1 nut
large long 1 key
small compact none screw
small compact 1 nut
large long none pen
large long 2 scissors
large long 1 pen
large other 2 scissors
small other 2 key
```

4.1.2 C4.5 format

This format is a derivative of the format that is used by the well-known C4.5 decision tree learning program [21]. The separator between the features is a comma, and the category (viz. the last feature on the line) is followed by a period (although this is not mandatory: TiMBL is robust to missing periods)[1]. White space within the line is taken literally, so the pattern `a, b c,d` will be interpreted as `'a',' b c','d'`. When using this format, especially with linguistic data sets or with data sets containing floating point numbers, one should take special care that commas do not occur in the feature-values and that periods do not occur within the category. Note that TiMBL's C4.5 format does not require a so called *namesfile*. However, TiMBL can produce such a file for C4.5 with the `-n` option. The C4.5 format is auto-detected when a comma is detected on the first line *before any white space has been encountered*. The example data set looks like this in the C4.5 format:

```
small,compact,1,nut.
small,long,none,screw.
small,long,1,key.
small,compact,1,nut.
large,long,1,key.
small,compact,none,screw.
```

[1] The periods after the category are not reproduced in the output

```
small,compact,1,nut.
large,long,none,pen.
large,long,2,scissors.
large,long,1,pen.
large,other,2,scissors.
small,other,2,key.
```

4.1.3 ARFF format

ARFF is a format that is used by the WEKA machine learning workbench [17][2]. Although TiMBL at present does *not* entirely follow the ARFF specification, it still tries to do as well as it can in reading this format. In ARFF the actual data are preceded by a header with various types of information and interspersed with lines of comments (starting with %). The ARFF format is auto-detected when the first line starts with % or @. TiMBL ignores lines with ARFF comments and instructions, and starts reading data from after the `@data` statement until the end of the file. The feature-values are separated by commas, and white space is deleted entirely, so the pattern `a, b c,d` will be interpreted as `'a','bc','d'`. We plan to include better support for the ARFF format in future releases.

```
% There are 4 attributes.
% There are 12 instances.
% Attribute information:                          Ints Reals  Enum  Miss
%          'size'                                    0     0    12     0
%          'shape'                                   0     0    12     0
%          'n_holes'                                 9     0     3     0
%          'class.'                                  0     0    12     0
@relation 'example.data'
@attribute 'size' { small, large}
@attribute 'shape' { compact, long, other}
@attribute 'n_holes' { 1, none, 2}
@attribute 'class.' { nut., screw., key., pen., scissors.}
@data
small,compact,1,nut.
small,long,none,screw.
small,long,1,key.
small,compact,1,nut.
large,long,1,key.
small,compact,none,screw.
small,compact,1,nut.
large,long,none,pen.
large,long,2,scissors.
large,long,1,pen.
large,other,2,scissors.
```

[2]WEKA is available from the Waikato University Department of Computer Science, `http://www.cs.waikato.ac.nz/~ml/`.

```
small,other,2,key.
```

4.1.4 Compact format

The compact format is especially useful when dealing with very large data files. Because this format does not use any feature separators, file-size is reduced considerably in some cases. The price of this is that all features and class labels must be of equal length (in characters) and TiMBL needs to know beforehand what this length is. You must tell TiMBL by using the `-l` option. The compact format is auto-detected when neither of the other formats applies. The same example data set might look like this in the column format (with two characters per feature):

```
smco1_nu
smlonosc
smlo1_ke
smco1_nu
lalo1_ke
smconosc
smco1_nu
lalonope
lalo2_sc
lalo1_pe
laot2_sc
smot2_ke
```

4.2 Weight files

The feature weights that are used for computing similarities and for the internal organization of the memory-base can be saved to a file. A file with weights can be constructed or altered manually and then read back into TiMBL. The format for the weights file is as follows. The weights file may contain comments on lines that start with a # character. The other lines contain the number of the feature followed by its numeric weight. An example of such a file is provided below. The numbering of the weights starts with 1 and follows the same order as in the data file. If features are to be ignored it is advisable not to set them to zero, but give them the value "`Ignored`" or to use the `-s` option.

```
# DB Entropy: 2.29248
# Classes: 5
# Lines of data: 12
# Fea.  Weight
1       0.765709
2       0.614222
3       0.73584
```

4.3 Tree files

Although the learning phase in TiMBL is relatively fast, it can sometimes be useful to store the internal representation of the data set for even faster subsequent retrieval. In TiMBL, the data set is stored internally in a tree structure (see Section 3.2). When using MBL, this tree representation contains all the training cases as full paths in the tree. When using IGTREE, unambiguous paths in the tree are pruned before it is used for classification or written to file. In either tree, the arcs represent feature-values and nodes contain class (frequency distribution) information. The features are in the same order throughout the tree. This order is either determined by memory-size considerations in MBL, or by feature relevance in IGTREE. It can explicitly be manipulated using the `-T` option.

We strongly advise to refrain from manually editing the tree file. However, the syntax of the tree file is as follows. After a header consisting of information about the algorithm and the feature-ordering (the permutation from the order in the data file to the order in the tree)[3] the tree's nodes and arcs are given in non-indented bracket notation.

Starting from the root node, each node is denoted by an opening parenthesis "(", followed by a default class. After this, there is an optional class distribution list, within curly braces "{ }", containing a non-empty list of categories followed by integer counts. After this comes an optional list of children, within "[ ]" brackets, containing a non-empty list of nodes. The choice whether distributions are present is maintained throughout the whole tree. Whether children are present is really dependent on whether children *are* present.

The MBL tree that was constructed from our example data set looks as follows:

```
# Algorithm: MBL
# Permutation: < 1, 3, 2 >
#
( nut { nut 3 screw 2 key 3 pen 2 scissors 2 }
 [ small ( nut { nut 3 screw 2 key 2 }
 [ 1 ( nut { nut 3 key 1 }
 [ compact ( nut { nut 3 }
 )
long ( key { key 1 }
 )
 ]
 )
none ( screw { screw 2 }
 [ compact ( screw { screw 1 }
 )
long ( screw { screw 1 }
```

[3] Although in this header each line starts with '#', these lines cannot be seen as comment lines.

```
 )
 ]
 )
2 ( key { key 1 }
 [ other ( key { key 1 }
 )
 ]
 )
 ]
 )
large ( pen { key 1 pen 2 scissors 2 }
 [ 1 ( key { key 1 pen 1 }
 [ long ( key { key 1 pen 1 }
 )
 ]
 )
none ( pen { pen 1 }
 [ long ( pen { pen 1 }
 )
 ]
 )
2 ( scissors { scissors 2 }
 [ long ( scissors { scissors 1 }
 )
other ( scissors { scissors 1 }
 )
 ]
 )
 ]
 )
 ]
 )
```

The corresponding compressed IGTREE version is much smaller. Note also that it does not contain a distribution, while an MBL tree must *always* contain distributions:

```
# Algorithm: IG-tree
# Permutation: < 1, 3, 2 >
#
( nut [ small ( nut [ 1 ( nut [ long ( key )
 ]
 )
none ( screw )
2 ( key )
 ]
```

```
 )
large ( pen [ 1 ( key )
2 ( scissors )
 ]
 )
 ]
 )
```

Chapter 5

Command line options

The user interacts with TiMBL through the use of command line arguments. When you have installed TiMBL successfully, and you type `Timbl` at the command line without any further arguments, it will print an overview of the most basic command line options.

```
TiMBL Version 1.0, (c) ILK 1998.
Tilburg Memory Based Learner
Induction of Linguistic Knowledge Research Group, Tilburg University.

usage:  Timbl -f data-file {-t test-file}
or see: Timbl -h
        for all possible options
```

If you are satisfied with all of the default settings, you can proceed with just these basics:

`-f <datafile>` : supplies the name of the file with the training items.

`-t <testfile>` : supplies the name of the file with the test items.

`-h` : prints a glossary of all available command line options.

The presence of a training file will make TiMBL pass through the first two phases of its cycle. In the first phase it examines the contents of the training file, and computes a number of statistics on it (feature weights etc.). In the second phase the instances from the training file are stored in memory. If no test file is specified, the program exits, possibly writing some of the results of learning to files (see below). If there is a test file, the selected classifier, trained on the present training data, is applied to it, and the results are written to a file of which name is a combination of the name of the test file and a code representing the chosen algorithm settings. TiMBL then reports the percentage of correctly classified test items. The default settings for the classification phase are: a Memory-Based Learner, with Gain Ratio feature weighting, with $k = 1$,

and with optimizations for speedy search. If you need to change the settings, because you want to use a different type of classifier, or because you need to make a trade-off between speed and memory-use, then you can use the options that are shown using `-h`. The sections below provide a reference to the use of these command line arguments, and they are roughly ordered by the type of action that the option has effect on.

5.1 Algorithm and Metric selection

`-a <n>` : chooses between the standard MBL (nearest neighbor search) algorithm (n=0, this is the default value), and the decision tree-based optimization IGTREE (n=1).

`-m <n>` : chooses between similarity metrics. Only applicable in conjunction with MBL (`-a 0`). The possible values are:

- n=0 – Weighted Overlap metric (default). See section 3.1.1. The difference between two feature-values is 1 if they are different and 0 if they are exactly the same. Can be used in combination with feature-weights that are specified using the `-w` argument.
- n=1 – Modified Value Difference Metric. See section 3.1.3. The difference between two feature-values is a continuous measure which depends on the difference between their conditional probability distribution over the target categories. The differences between all pairs of feature-values are computed before the test phase, unless the number of feature-values is too large, or the `--` option is used. This metric can be used in combination with feature-weights that are specified using the `-w` argument.

`-w <n>` : chooses between feature-weighting possibilities. The weights are used in the metric of MBL and in the ordering of the IGTREE. Possible values are:

- n=0 – No weighting, i.e. all features have the same importance (weight = 1).
- n=1 – Gain Ratio weighting (default). See section 3.1.2.
- n=2 – Information Gain weighting. See section 3.1.2.
- n=filename – Instead of a number we can supply a filename to the `-w` option. This causes TiMBL to read this file and use its contents as weights. (See section 4.2 for a description of the weights file)

`-k <n>` : Number of nearest neighbors used for extrapolation. Only applicable in conjunction with MBL (`-a 0`). The default is 1. Especially with the MVDM metric it is often useful to determine a good value larger than 1 for this parameter (usually an odd number, to avoid ties). Note that due to

ties (instances with exactly the same similarity to the test instance) the number of instances used to extrapolate might in fact be much larger than this parameter.

`-R <n>` : Resolve ties in the classifier randomly, using a random generator with seed n. As a default this is OFF, and ties are resolved in favor of the category which is more frequent in the training set as a whole—remaining ties are resolved on a first come first served basis.

`-t <@file>` : If the filename given after `-t` starts with '`@`', TiMBL will read commands for testing from `file`. This file should contain one set of instructions per line. On each line new values can be set for the following command line options: `-d -D -e -k -m -o -O -p -P -R -v -w -x -% --`. It is compulsory that each line contains a `-t <file>` argument to specify the name of the test file.

5.2 Input options

`-F <format>` : Force TiMBL to interpret the training and test file as a specific data format. Possible values for this parameter are: `Compact, C4.5, ARFF, Columns` (case-sensitive). The default is that TiMBL guesses the format from the contents of the first line of the data file. See section 4.1 for description of the data formats and the guessing rules.

`-s <n,...>` : Skip features n,... After the `-s` option a string is given with a comma-separated list of features which will be ignored during training and testing. The effect is the same as setting a feature's weight to the value `Ignored`. This has an advantage over setting the weights to zero, because zero-weighted features are still present in the learner's internal representation and can have undesirable side-effects, especially with the IGTREE algorithm.

`-l <n>` : Feature length. Only applicable with the Compact data format; `<n>` is the number of characters used for each feature-value and category symbol.

`-i <treefile>` : Skip the first two training phases, and instead of processing a training file, read a previously saved (see `-I` option) instance-base or IGTREE from the file `treefile`. See section 4.3 for the format of this file.

`-P <path>` : Specify a path to read the data files from. This path is ignored if the name of the data file already contains path information.

5.3 Output options

`-I <treefile>` : After phase one and two of learning, save the resulting tree-based representation of the instance-base or IGTREE in a file. This file can

later be read back in using the `-i` option (see above). See section 4.3 for a description of the resulting file's format.

`-d` : Keep distributions. This option only has effect with the above `-I` option, and causes the information about target category frequencies to be retained in the tree file. With MBL this is always ON. For IGTREE it is OFF; turning it ON has no effects on classification accuracy, but only writes the distributions in the tree file.

`-W <file>` : Save the used feature-weights in a file.

`-n <file>` : Save the feature-value and target category symbols in a C4.5 style "names file" with the name `<file>`.

`-p <n>` : Indicate progress during training and testing after every n processed patterns. The default setting is 10000.

`-e <n>` : During testing, compute and print an estimate on how long it will take to classify n test patterns. This is off by default.

`-v <n>` : Verbosity Level; determines how much information is written to standard output during a run. This parameter can take on the following values:

n=0 – output just the minimal amount of information.

n=1 – give an overview of the settings.

n=2 – show the computed feature weights (this is the default)

n=8 – show each exact match,

Setting n to be the sum of any number of the above values, results in combined levels of verbosity.

`-D` : Write the distance of the nearest neighbor of each test item. In the case of the IGTREE algorithm the resulting number represents the depth of the tree at which the classification decision was made.

`-%` : Write the percentage of correctly classified test instances to a file with the same name as the output file, but with the suffix "`.%`".

`-o <suffix>` : Add `suffix` to the name of the output file. Useful for different runs with the same settings on the same testfile.

`-O <path>` : Write all output to the path given here. The default is to write all output to the directory where the test file is located.

5.4 Internal representation options

`-T <n>` : Order the instance-base according to one of the following measures. Different measures produce different tree sizes, and thus this option can be used to get smaller memory usage, depending on the peculiarities of the data set

n=1 – use the order of the features in the training file.

n=2 – use Gain Ratio to order features (default for IGTREE).

n=3 – use Information Gain to order the features.

n=4 – order according to the quantity $\frac{1}{number\ of\ feature\ values}$.

n=5 – order according to the quantity $\frac{GainRatio}{number\ of\ feature\ values}$. (default for MBL)

n=6 – order according to the quantity $\frac{InformationGain}{number\ of\ feature\ values}$.

`-x` : Turns off the shortcut search for exact matches in MBL. The default is for this to be ON (which is usually much faster), but when $k > 1$, the shortcut produces different results from a "real" k nearest neighbors search.

`--` : Turn off the use of "memory-for-speed" optimizations. This option has a different effect depending on which metric is used. With the Weighted Overlap metric (`-m 0`), it turns off the computation of inverted files. Turning this off will make testing slower, but reduces the memory load approximately by a half. With the MVDM metric, (`-m 1`) it turns off the precomputation of the value difference matrices. Turning this off will make testing slower, but is sometimes a sheer necessity memory-wise. With both metrics, the default is ON.

Bibliography

[1] D. W. Aha. Lazy learning: Special issue editorial. *Artificial Intelligence Review*, 11:7–10, 1997.

[2] D. W. Aha, D. Kibler, and M. Albert. Instance-based learning algorithms. *Machine Learning*, 6:37–66, 1991.

[3] R. H. Baayen, R. Piepenbrock, and H. van Rijn. *The CELEX lexical data base on CD-ROM.* Linguistic Data Consortium, Philadelphia, PA, 1993.

[4] C. Cardie. Automatic feature set selection for case-based learning of linguistic knowledge. In *Proc. of Conference on Empirical Methods in NLP.* University of Pennsylvania, 1996.

[5] S. Cost and S. Salzberg. A weighted nearest neighbour algorithm for learning with symbolic features. *Machine Learning*, 10:57–78, 1993.

[6] T. M. Cover and P. E. Hart. Nearest neighbor pattern classification. *Institute of Electrical and Electronics Engineers Transactions on Information Theory*, 13:21–27, 1967.

[7] W. Daelemans. Memory-based lexical acquisition and processing. In P. Steffens, editor, *Machine Translation and the Lexicon*, volume 898 of *Lecture Notes in Artificial Intelligence*, pages 85–98. Springer-Verlag, Berlin, 1995.

[8] W. Daelemans and A. Van den Bosch. Generalisation performance of backpropagation learning on a syllabification task. In M. F. J. Drossaers and A. Nijholt, editors, *Proc. of TWLT3: Connectionism and Natural Language Processing*, pages 27–37, Enschede, 1992. Twente University.

[9] W. Daelemans, S. Gillis, and G. Durieux. The acquisition of stress: a data-oriented approach. *Computational Linguistics*, 20(3):421–451, 1994.

[10] W. Daelemans and A. Van den Bosch. Tabtalk: reusability in data-oriented grapheme-to-phoneme conversion. In *Proceedings of Eurospeech '93*, pages 1459–1466, Berlin, 1993. T.U. Berlin.

[11] W. Daelemans and A. Van den Bosch. Language-independent data-oriented grapheme-to-phoneme conversion. In J. P. H. Van Santen, R. W. Sproat,

J. P. Olive, and J. Hirschberg, editors, *Progress in Speech Processing*, pages 77–89. Springer-Verlag, Berlin, 1996.

[12] W. Daelemans, J. Zavrel, P. Berck, and S. Gillis. MBT: A memory-based part of speech tagger generator. In E. Ejerhed and I.Dagan, editors, *Proc. of Fourth Workshop on Very Large Corpora*, pages 14–27. ACL SIGDAT, 1996.

[13] W. Daelemans, A. Van den Bosch, and A. Weijters. IGTree: using trees for compression and classification in lazy learning algorithms. *Artificial Intelligence Review*, 11:407–423, 1997.

[14] W. Daelemans, A. Van den Bosch, and J. Zavrel. A feature-relevance heuristic for indexing and compressing large case bases. In M. Van Someren and G. Widmer, editors, *Poster Papers of the Ninth European Conference on Machine Learing*, pages 29–38, Prague, Czech Republic, 1997. University of Economics.

[15] W. Daelemans, P. Berck, and S. Gillis. Data mining as a method for linguistic analysis: Dutch diminutives. *Folia Linguistica*, XXXI(1-2), 1997.

[16] P.A. Devijver and J. Kittler. *Pattern recognition. A statistical approach.* Prentice-Hall, London, UK, 1982.

[17] S.R. Garner. WEKA: The Waikato Environment for Knowledge Analysis. In *Proc. of the New Zealand Computer Science Research Students Conference*, pages 57–64, 1995.

[18] S. Gillis, G. Durieux, and W. Daelemans. A computational model of P&P: Dresher and kaye (1990) revisited. In M. Verrips and F. Wijnen, editors, *Approaches to parameter setting*, volume 4 of *Amsterdam Studies in Child Langage Development*, pages 135–173. 1995.

[19] J. Kolodner. *Case-based reasoning.* Morgan Kaufmann, San Mateo, CA, 1993.

[20] J.R. Quinlan. Induction of Decision Trees. *Machine Learning*, 1:81–206, 1986.

[21] J.R. Quinlan. C4.5*: Programs for Machine Learning.* Morgan Kaufmann, San Mateo, CA, 1993.

[22] C. Stanfill and D. Waltz. Toward memory-based reasoning. *Communications of the* ACM, 29(12):1213–1228, December 1986.

[23] A. Van den Bosch. *Learning to pronounce written words: A study in inductive language learning.* PhD thesis, Universiteit Maastricht, 1997.

[24] A. Van den Bosch and W. Daelemans. Data-oriented methods for grapheme-to-phoneme conversion. In *Proceedings of the 6th Conference of the EACL*, pages 45–53, 1993.

[25] A. Van den Bosch, W. Daelemans, and A. Weijters. Morphological analysis as classification: an inductive-learning approach. In K. Oflazer and H. Somers, editors, *Proceedings of the Second International Conference on New Methods in Natural Language Processing, NeMLaP-2, Ankara, Turkey*, pages 79–89, 1996.

[26] S. Weiss and C. Kulikowski. *Computer systems that learn.* San Mateo, CA: Morgan Kaufmann, 1991.

[27] J. Zavrel and W. Daelemans. Memory-based learning: Using similarity for smoothing. In *Proc. of 35th annual meeting of the ACL*, Madrid, 1997.

[28] J. Zavrel, W. Daelemans, and J. Veenstra. Resolving PP-attachment ambiguities with memory-based learning. In Mark Ellison, editor, *Proc. of the Workshop on Computational Natural Language Learning (CoNLL'97), ACL*, Madrid, 1997.

Appendix A

Tutorial: a case study

In this tutorial is meant to get you started with TiMBL quickly. We discuss how to format the data of a task to serve as training examples, which choices can be made during the construction of the classifier, how various choices can be evaluated in terms of their generalization accuracy, and various other practical issues. The reader who is interested in more background information on TiMBL implementation issues and a formal description of Memory-Based Learning, is advised to read Chapter 3.

Memory-Based Learning (MBL) is based on the idea that intelligent behavior can be obtained by analogical reasoning, rather than by the application of abstract *mental rules* as in rule induction and rule-based processing. In particular, MBL is founded in the hypothesis that the extrapolation of behavior from stored representations of earlier experience to new situations, based on the similarity of the old and the new situation, is of key importance.

MBL algorithms take a set of examples (fixed-length patterns of feature-values and their associated class) as input, and produce a *classifier* which can classify new, previously unseen, input patterns. Although TiMBL was designed with linguistic classification tasks in mind, it can in principle be applied to any kind of classification task with discrete features and categories for which training data is available. The only limitation is that numeric features are at present not supported, and will be treated as unordered discrete values. As an example task for this tutorial we go through the application of TiMBL to the prediction of Dutch diminutive suffixes. The necessary data sets are included in the TiMBL distribution, so you can replicate the examples given below on your own system.

A.1 Data

The operation of TiMBL will be illustrated below by means of a real natural language processing task: prediction of the diminutive suffix form in Dutch [15]. In Dutch, a noun can receive a diminutive suffix to indicate *small size* literally

or metaphorically attributed to the referent of the noun; e.g. *mannetje* means *little man*. Diminutives are formed by a productive morphological rule which attaches a form of the Germanic suffix *-tje* to the singular base form of a noun. The suffix shows variation in its form (Table A.1). The task we consider here is to predict which suffix form is chosen for previously unseen nouns on the basis of their form.

Noun	Form	Suffix
huis (house)	huisje	*-je*
man (man)	mannetje	*-etje*
raam (window)	raampje	*-pje*
woning (house)	woninkje	*-kje*
baan (job)	baantje	*-tje*

Table A.1: Allomorphic variation in Dutch diminutives.

For these experiments, we collect a representation of nouns in terms of their syllable structure as training material[1]. For each of the last three syllables of the noun, four different features are collected: whether the syllable is stressed or not (values - or +), the string of consonants before the vocalic part of the syllable (i.e. its onset), its vocalic part (nucleus), and its post-vocalic part (coda). Whenever a feature value is not present (e.g. a syllable does not have an onset, or the noun has less than three syllables), the value '=' is used. The class to be predicted is either E (-etje), T (-tje), J (-je), K (-kje), or P (-pje).

Some examples are given below (the word itself is only provided for convenience and is not used). The values of the syllabic content features are given in phonetic notation.

-	b	i	=	-	z	@	=	+	m	A	nt	J	biezenmand
=	=	=	=	=	=	=	=	+	b	I	x	E	big
=	=	=	=	+	b	K	=	-	b	a	n	T	bijbaan
=	=	=	=	+	b	K	=	-	b	@	l	T	bijbel

Our goal is to use TiMBL in order to train a classifier that can predict the class of new, previously unseen words as correctly as possible, given a set of training examples that are described by the features given above. Because the basis of classification in TiMBL is the storage of all training examples in memory, a test of the classifier's accuracy must be done on a separate test set. We will call these datasets `dimin.train` and `dimin.test`, respectively. The training set `dimin.train` contains 3000 words and the test set contains 950 words, none of which are present in the training set. Although a single train/test partition suffices here for the purposes of explanation, it does not factor out the bias of choosing this particular split. Unless the test set is sufficiently large, a more reliable generalization accuracy measurement is used in real experiments,

[1] These words were collected form the CELEX lexical database [3]

e.g. 10-fold cross-validation [26]. This means that 10 separate experiments are performed, and in each "fold" 90% of the data is used for training and 10% for testing, in such a way that each instance is used as a test item exactly once.

A.2 Using TiMBL

Different formats are allowed for training and test data files. TiMBL is able to guess the type of format in most cases. We will use comma-separated values here, with the class as the last value. This format is called C4.5 format in TiMBL because it is the same as that used in Quinlan's well-known C4.5 program for induction of decision trees [21]. See Section 4 for more information about this and other file formats.

An experiment is started by executing TiMBL with the two files (`dimin.train` and `dimin.test`) as arguments:

```
Timbl -f dimin.train -t dimin.test
```

Upon completion, a new file has been created with name `dimin.test.mbl.wo.gr.k1.out`, which is in essence identical to the input test file, except that an extra comma-separated column is added with the class predicted by TiMBL. The name of the file provides information about the MBL algorithms and metrics used in the experiment (the default values in this case). We will describe these shortly.

Apart from the result file, information about the operation of the algorithm is also sent to the standard output. It is therefore advisable to redirect the output to a file in order to make a log of the results.

```
Timbl -f dimin.train -t dimin.test > dimin-exp1
```

The defaults used in this case work reasonably well for most problems. We will now provide a point by point explanation of what goes on in the output.

```
TiMBL Version 1.0, (c) ILK 1998.
Tilburg Memory Based Learner
Induction of Linguistic Knowledge Research Group, Tilburg University
Wed Mar 11 14:38:35 1998

Examine datafile gave the following results:
Number of Features: 12
InputFormat       : C4.5
```

TiMBL has detected 12 features and the C4.5 input format (comma-separated features, class at the end).

```
Phase 1: Reading Datafile: dimin.train
Start:          0 @ Wed Mar 11 14:38:35 1998
Finished:    2999 @ Wed Mar 11 14:38:36 1998

Calculating Entropy       Wed Mar 11 14:38:36 1998
Lines of data     : 2999
DB Entropy        : 1.6178929
Number of Classes : 5

Feature  Values   SplitInfo     InfoGain        GainRatio
    1         3   1.2442408     0.030971064     0.024891536
    2        50   2.2089001     0.060860038     0.027552191
    3        19   2.1182903     0.039562857     0.018676787
    4        37   0.99848875    0.052541227     0.052620750
    5         3   1.5623570     0.074523225     0.047699231
    6        61   4.3333080     0.10604433      0.024471911
    7        20   3.5329144     0.12348668      0.034953203
    8        69   2.2098731     0.097198760     0.043983864
    9         2   0.97726616    0.045752381     0.046816705
   10        64   4.9921730     0.21388759      0.042844587
   11        18   3.6186520     0.66970458      0.18507018
   12        43   3.9280484     1.2780762       0.32537181

Feature Permutation based on GainRatio/Values :
< 9, 5, 11, 1, 12, 7, 4, 3, 10, 8, 2, 6 >
```

Phase 1 is the training data analysis phase. Time stamps for start and end of analysis are provided. Some preliminary analysis of the training data is done: number of training items, number of classes, entropy of the training data. For each feature, the number of values, and three variants of an information-theoretic measure of feature relevance are given. These are used both for memory organization during training and for feature relevance weighting during testing (see Chapter 3). Finally, an ordering (permutation) of the features in terms of decreasing relevance for solving the task is provided.

```
Phase 2: Learning from Datafile: dimin.train
Start:          0 @ Wed Mar 11 14:38:36 1998
Finished:    2999 @ Wed Mar 11 14:38:36 1998
```

Phase 2 is the learning phase; all training items are stored in an efficient way in memory for use during testing. Again timing (real time) is provided.

```
Phase 3: Starting to test, Testfile: dimin.test
Algorithm   : MBL
Test metric : weighted overlap (Using Inverted files, preferring exact matches)
Weighting   : GainRatio

Calculating inverted files
Writing output in: ./dimin.test.mbl.wo.gr.k1.out
Tested:      1 @ Wed Mar 11 14:38:37 1998
Tested:      2 @ Wed Mar 11 14:38:37 1998
Tested:      3 @ Wed Mar 11 14:38:37 1998
Tested:      4 @ Wed Mar 11 14:38:37 1998
Tested:      5 @ Wed Mar 11 14:38:37 1998
Tested:      6 @ Wed Mar 11 14:38:37 1998
Tested:      7 @ Wed Mar 11 14:38:37 1998
Tested:      8 @ Wed Mar 11 14:38:37 1998
Tested:      9 @ Wed Mar 11 14:38:37 1998
Tested:     10 @ Wed Mar 11 14:38:37 1998
Ready:     950 @ Wed Mar 11 14:39:04 1998
Seconds taken: 27 (35.19 p/s)
918/950 (0.966316), of which 39 exact matches
```

In Phase 3, the trained classifier is applied to the test set. Because we have not specified which algorithm to use, the default settings are used (MBL with information theoretic feature weighting). This algorithm computes the similarity between a test item and each training item in terms of *weighted overlap*: the total difference between two patterns is the sum of the relevance weights of those features which are not equal. The class for the test item is decided on the basis of the least distant item(s) in memory. To compute relevance, Gain Ratio is used (an information-theoretic measure, see Section 3.1.2). Time stamps indicate the progress of the testing phase. Finally, accuracy on the test set is logged, and the number of exact matches[2]. In this experiment, the diminutive suffix form of 96.6% of the new words was correctly predicted.

The meaning of the output file names can be explained now:
`dimin.test.mbl.wo.gr.k1.out` means output file (`.out`) for `dimin.test` with algorithm MBL, similarity computed as *weighted overlap* (`.wo`), relevance weights computed with *gain ratio* (`.gr`), and number of most similar memory patterns on which the output class was based equal to 1 (`.k1`).

A.3 Algorithms and Metrics

A precise discussion of the different algorithms and metrics implemented in TiMBL is given in Chapter 3. We will discuss the effect of the most important

[2] An exact match in this experiment can occur when two different nouns have the same feature-value representation.

ones on our data set.

A first choice in algorithms is between using MBL and IGTREE. In the trade-off between generalization accuracy and efficiency, MBL usually, but not always, leads to more accuracy at the cost of more memory and slower computation, whereas IGTREE is a fast heuristic approximation of MBL, but sometimes less accurate. The IGTREE algorithm is used when `-a 1` is given on the command line, whereas the MBL algorithm used above (the default) would have been specified explicitly by `-a 0`.

```
Timbl -a 1 -f dimin.train -t dimin.test
```

When using the MBL algorithm, there is a choice of metrics for influencing the definition of similarity. With *weighted overlap*, each feature is assigned a weight, determining its relevance in solving the task. With the *modified value difference metric* (MVDM), each pair of values of a particular feature is assigned a value difference. The intuition here is that in our diminutive problem, for example, the codas n and m should be regarded as being more similar than n and p. These pair-wise differences are computed for each pair of values in each feature (see Section 3.1.3). Selection between weighted overlap and MVDM is done by means of the `-m` parameter. The following selects MVDM, whereas `-m 0` (*weighted overlap*) is the default.

```
Timbl -m 1 -f dimin.train -t dimin.test
```

Especially when using MVDM, but also in other cases, it may be useful to extrapolate not just from the most similar example in memory, which is the default, but from several. This can be achieved by using the $-k$ parameter followed by the wanted number of nearest neighbors. E.g., the following applies MBL with the MVDM metric, with extrapolation from the 5 nearest neighbors.

```
Timbl -m 1 -k 5 -f dimin.train -t dimin.test
```

Within the MBL *weighted overlap* option, the default feature weighting method is Gain Ratio. By setting the parameter `-w` to 0, an *overlap* definition of similarity is created where each feature is considered equally relevant. Similarity reduces in that case to the number of equal values in the same position in the two patterns being compared. As an alternative weighting, users can provide their own weights by using the `-w` parameter with a filename in which the feature weights are stored (see Section 4.2 for a description of the format of the weights file).

Table A.2 shows the effect of algorithm, metric, and weighting method choice on generalization accuracy for our training - test set partition. We see that IGTREE performs slightly worse than MBL for this task (it uses less memory and is faster, however). When comparing MVDM and *feature weighting*, we see that the overall best results are achieved with MVDM, but only with a relatively high value for k, the number of memory items on which the extrapolation is based. Increasing the value of k for (weighted) Overlap metrics decreased performance.

	gain ratio	inform. gain	overlap	MVDM
IGTREE	96.4	96.4		
MBL, $-k1$	96.6	96.5	84.9	96.2
MBL, $-k10$				97.8

Table A.2: Some results for diminutive prediction.

Within the feature weighting approaches, overlap (i.e. no weighting) performs markedly worse than the default *information gain* or *gain ratio* weighting methods.

A.4 More Options

Several input and output options exist to make life easier while experimenting. See Chapter 5 for a detailed description of these options. One especially useful option for testing linguistic hypotheses is the `-s` command line option, which allows you to skip certain features when computing similarity. E.g. if we want to test the hypothesis that only the rime (nucleus and coda) of the last syllable are actually relevant in determining the form of the diminutive suffix, we can execute the following to disregard all but the last two features. As a result we get an accuracy of 97.4%[3].

```
Timbl -s 1,2,3,4,5,6,7,8,9,10 -f dimin.train -t dimin.test
```

The last parameter we discuss here is the `-D` command line option which has as effect that in the output file not only the extrapolated class is appended to the input pattern, but also the distance to the nearest neighbor.

```
Timbl -D -f dimin.train -t dimin.test
```

The resulting output file contains lines like the following.

```
-,t,@,=,-,l,|,=,-,G,@,n,T,T          0.099723
-,=,I,n,-,str,y,=,+,m,E,nt,J,J          0.123322
=,=,=,=,=,=,=,=,+,br,L,t,J,J          0.042845
=,=,=,=,+,zw,A,=,-,m,@,r,T,T          0.059425
=,=,=,=,-,f,u,=,+,dr,a,l,T,T          0.077798
=,=,=,=,=,=,=,=,+,l,e,w,T,T          0.042845
=,=,=,=,+,tr,K,N,-,k,a,rt,J,J          0.068456
=,=,=,=,+,=,o,=,-,p,u,=,T,T          0.172314
=,=,=,=,=,=,=,=,+,l,A,m,E,E          0.042845
=,=,=,=,=,=,=,=,+,l,A,p,J,J          0.042845
=,=,=,=,=,=,=,=,+,sx,E,lm,P,P          0.042845
+,l,a,=,-,d,@,=,-,k,A,st,J,J          0.070701
```

[3]It should be kept in mind that the amount of overlap in training and test set has significantly increased, so that generalization is based on retrieval more than on similarity computation.

This can be used to study how specific instances (low distance) and more general patterns (higher distance) are used in the process of generalization.

Summarizing, we hope that this tutorial has made it clear that, once you have coded your data in fixed-length discrete feature-value patterns, it should be relatively straightforward to get the first results using TiMBL. You can then experiment with different metrics and algorithms to try and further improve your results. We wish you happy learning!