Proceedings of the Institute of Acoustics

SYLLABLE BASED HYPOTHESIS REFINEMENT IN SYLK

Luke Boucher, P.D.Green,
SPLASH (Speech Laboratory), University of Sheffield, UK

ABSTRACT

A framework for hypothesis refinement in the SYLK project is described. For a general overview of SYLK see the companion paper in this volume, (Green et al[3]). In SYLK, statistical and knowledge based techniques are used to construct a model of the syllable upon which evidential reasoning can be applied for classification. Both Bayesean updating and Shafer's Belief Functions are considered as evidential reasoning formalisms which satisfy the structure and constraints implied by the model and a form of constrained Dempster's rule is suggested.

INTRODUCTION

SYLK is (a project which aims to produce) an ASR front end in which classification takes place around the Syllable (Green et al [3]). Given a rough, mid class, segmentation of some utterance - of which only the approximate location of the Syllable nuclei is really essential - SYLK attempts to classify each syllable independently as an ordered lattice consisting of possible syllable onsets, peaks and codas. A Statistical and Knowledge-based model of the syllable is used.

In what follows we shall consider systems of hypothesis refinement which can be used to classify evidence within the constraints of this model. In the next section, then, we sketch out the model's characteristics and define what is required of hypothesis refinement. After this, Bayesean updating (Dempster[2]) and Belief Functions (Shafer [6]) are compared, with respect to one of the constraints, and a method is proposed. Finally, ways of satisfying the remaining constraints within this chosen formalism are suggested

THE CONSTRAINTS *(the syllable model)*

The syllable model is a network structure of nodes and arcs with both constituent and refinement planes. Thus, a Syllable has constituents, ordered in time, optional onset followed by rhyme and a rhyme has constituents Peak followed by Optional Coda, each of which must be classified. This classification is done in terms of probabilistic decisions made down a refinement network associated with each of these syllable constituents. Output from the refinement network is a lattice consisting of the deepest nodes reached ordered in terms of their certainty value. (we use the word "certainty" here to emphasise that the formalism need not use probabilities, even though probabilities have been adopted for SYLK). The rhyme refinements, however, do not make part of the output lattice. Instead, they are used to model known dependences which exist between the peak and coda by constraining the co-occurence of various peak and coda refinements. The refinements of rhyme, then, will have refinements of peak and coda as their constituents.

These probabilistic decisions are made with the use of refinement tests. Tests, are located at refinement

nodes (a test may be linked with one or several nodes) and are said to have a "domain", which is the set of most basic syllable constituents whose ancestor is one of those nodes. A test has a "process", which examines some representation of the evidence to return a vector, and a statistical description of the performance of various subsets of its domain - generally, those syllable constituents reached by the arcs of the node(s) at which it applies - over the process. Given some new evidence, then, a Bayesian classifier can be used to distinguish between these sets.

To complete hypothesis refinement, then, a method of evidential reasoning is needed to mediate the certainties returned from many different tests, accommodating the constraints between different refinement planes, down towards the end nodes. This evidential reasoning formalism should;

(A) compliment the network structure of our refinement planes; eg, syllable constituents should only be considered as hypotheses, with measures of certainty, at a level as deep in the network (as specific) as the present state of evidence has suggested.

(B) cope with the constraints that exist between constituent refinement planes, where, for example, refining the peak may affect refinements of the coda.

(C) be made to account for the non independence which exists between the different tests.

THE ALTERNATIVES *(evidential reasoning formalisms)*

In this section, I shall consider both Bayesian Updating (Dempster[2]) and Belief Functions (Shafer[6]), assuming that the reader has a knowledge of Bayes and some acquaintance with Belief Functions. Initially, this will be a general discussion, relating mainly to constraint (A) from the last section. Constraints (B) and (C) will then be considered in the following section on extension to the chosen formalism, as they are not considered as relevant to its choice as (A).

Traditionally, Bayesian updating has been used for evidential reasoning, but more recently people (perhaps computer scientists rather than statisticians) have criticised it (Shafer[8]), calling for a technique which; (i), does not require prior probabilities, and (ii), has two measures - of support (ie probability) and knowledge (or confidence in that support) - rather than the single measure of probability. These are both relevant to (A). Bayes is traditionally centered about singleton hypotheses, and this would seem problematic in that prior probabilities assigned to singleton hypotheses suggest a level of specificity that is not reflected by any evidence.

It is with respect to these issues that Shafer's Belief functions appear to offer some solution. Belief Functions are presented as an alternative to Bayes which, it is claimed (Shafer[7]), solve problems (i) and (ii). However, they too have been criticised as having no real semantics and thus no real justification. Their inventor, Shafer, has defended them against this charge (Shafer[8]), by saying that probability, itself, has only an anecdotal semantics, providing a similar anecdotal view of his alternatives to probability, Mass and Belief. In the space provided here we can not go into a full consideration of this semantics. However, we will present a brief description of Belief Functions including an example which reveals the problems that they have with respect to satisfying (i) and (ii).

With Belief Functions, unity of mass is distributed, unlike probability, over the power set of some mutually exclusive set of possibilities. This means that we can ignore the need for prior probabilities (criticism (i)) by

representing our state of no knowledge as a mass of 1 assigned to the whole hypothesis set. However, as we shall see, this is not such a useful abstraction as Shafer claims. It is possible to treat mass as probability. This is done by constructing exclusiveness between the members of the possibility power set. So that, for example;

$$S = \{\{a\ b\ c\}\{a\ b\}\{a\ c\}\{b\ c\}\{a\}\{b\}\{c\}\} \text{ becomes}$$

$$S = \{\{a_a\ b_a\ c_a\}\{a_b\ b_b\}\{a_c\ c_b\}\{b_c\ c_c\}\{a_d\}\{b_d\}\{c_d\}\}$$

$$\text{where } a = \{a_a\ a_b\ a_c\ a_d\} \text{ etc}$$

Given an assignment of mass over this new power set, we find that treating mass as probability we can calculate lower and upper bounds of probability for a, b & c, using (1), which are equivalent to their Belief and Plausibility as defined for Belief Functions. These two measures, of belief and Plausibility, have been claimed as the solution to problem (ii) in that we have a window of no-knowledge, rather than a probabilistic point. But if this window is simply an upper and lower bound on probability, then the difference between them is specificity, rather than a measure of knowledge.

$$p(a) = \sum_i (p(b_i).p(a \backslash b_i)) \qquad (1)$$

for every set, b_i, upon which a depends.

Criticism (ii) is not so relevant to our needs however, and an awareness of specificity would seem to suit (A) although we do not need to actually have a measure of it. The claim over prior probabilities is more important.

Belief functions use Dempster's rule, as an alternative to Bayesian updating (or rather, as a generalisation of it). It is used to combine the evidence from two independent sources. Consider this example;

Mass is assigned over the power set of $\{a\ b\ c\ d\}$ in two distributions;

Ma1		Ma2 is	
	$\{a\} = 0.3$		$\{a\} = 0.3$
	$\{b\} = 0.3$		$\{b\} = 0.3$
	$\{c,d\} = 0.4$		$\{c\} = 0.2$
			$\{d\} = 0.2$

with all the other possible sets assigned mass 0. First consider what is meant by these distributions, and what the differences are between their meanings. We might want to say that Ma1 is in a state of less specificity than Ma2, as the latter has the mass assigned to $\{c,d\}$ evenly distributed amongst its parts. However, we could say that both are in the same state of uncertainty.

Now consider using Dempster's rule to combine each of these with a third mass distribution;

Ma3 is	
	$\{a,c\} = 0.5$
	$\{b,d\} = 0.5$

This distribution says, effectively, that $\{a,c\}$ and $\{b,d\}$ are just as likely, but that nothing is known about the distribution of a as to c and b as to d.

To apply Dempster's rule every element of one mass assignment is intersected with every element of the other,

and a new mass assigned to these intersections as the product of the mass assigned to the intersecting elements. New masses assigned to the same sets are summed. With Ma1 and Ma3, then we get;

$$\{a\} = 0.5*0.3\ \{b\} = 0.3*0.5\ \{c\} = 0.5*0.4\ \{d\} = 0.5*0.4\ \{\} = 0.3*0.5+0.3*0.5$$

Note, that, the masses assigned to these intersections still sum to one. However, we have mass assigned to the empty set, which Shafer normalises out, by dividing each of the other masses by F, where F = 1 - mass({null}). Finally then, we can consider the combination of Ma1 with Ma3 and Ma2 with Ma3.

Ma1 o Ma3 = Ma4 and Ma2 o Ma3 = Ma5

Ma4 is		Ma5 is	
	$\{a\}$ = 0.2142857		$\{a\}$ = 0.3
	$\{b\}$ = 0.2142857		$\{b\}$ = 0.3
	$\{c\}$ = 0.28571427		$\{c\}$ = 0.2
	$\{d\}$ = 0.28571427		$\{d\}$ = 0.2

The two possible outcomes are very different. The former, Ma4, assigning more weight, on the whole, to the pair $\{c,d\}$, and the latter, Ma5, to $\{a,b\}$. This is problematic, in that neither of the primary distributions, Ma1 or Ma2 suggest that $\{c,d\}$ is assigned greater value than $\{a,b\}$, in fact both suggest that the reverse is the case. There is nothing to suggest this in Ma3 either. And yet Ma4 has made that deduction.

What has happened is that in combining Ma3 with Ma1, whilst calculating the new mass assignment to c, c has been given the benefit of the 'non specific information' doubt between d or c and assigned $p(c) = 0.4$ whilst $p(d) = 0$. At the same time, however, the benefit has also been given to d assuming that $p(d) = 0.4$ and $p(c) = 0$, which contradicts the previous assumption. These disproportionate weight assignments are then covered up during normalisation, at the expense of the remaining hypothesis sets, $\{a,b\}$, which, as they were assigned mass on the singleton level did not have the collective weight to keep a fair proportion of the new assignments to themselves.

From this view of the Belief Functional use of Dempster's rule, it would seem reasonable to suggest that the second resultant, Ma5 is the favourable one for both instances. However, suggesting that Ma2 is always used in place of Ma1 makes an assumption of specificity, assigning equal mass to c and d, effectively loosing the benefit Belief Functions afford us in requiring no prior probabilities (criticism (i)). We can, however, choose to make the assumption of specificity afforded by Ma2 over Ma1, (ie to have prior probabilities) only in those cases where the new evidence is going to require that we make such an assumption. This is trivial to implement in that we simply have to determine what new discriminations some piece of evidence is going to make and will prevent the problematic results shown in the example. However, we are left with the fact that this use of Dempster's rule, affords us only a gain in convenience over Bayes, which could equally well be implemented this way.

EXTENSIONS

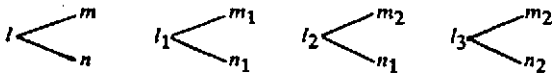
Given our choice of a constrained Dempster's rule we are now in a position to turn to constraints (B) and (C). As we have seen, issues surrounding the use of Bayes or Belief Functions have not really been concerned with (B) or (C). To this extent, then, our choice of formalism is not important as we will have to develop our own procedures which satisfy the constraints and work within the evidential reasoning without impairing it. Work

is still underway in formalising such procedures and will doubtless be the subject of future publications. A sketch, however, of the forms under consideration can be given.

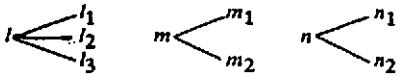
First, for a method of enforcing the constituent constraints of the model (B). Consider this example, removed from the domain of syllable hypotheses.

A unit l has constituents m and n, and refinements l_1, l_2 & l_3, acting as the Syllabic Rhyme to constrain co-refinements of m & n. Further to this; m has refinements m_1 & m_2, n has refinements n_1 & n_2 l_1 consists of m_1 & n_1, l_2 of m_2 & n_1, and l_3 of m_2 & n_2. This effectively bars the co-occurence of m_1 and n_2 and enables us to have separate statistical information about m_2 for both of its right contexts, n_1 or n_2.

ie, Constituent planes



and Refinement planes



Now, starting from a point of no knowledge;

$$p(l) = p(m) = p(n) = 1$$

Say we were to receive some evidence, from a test at l, suggesting that;

$$l_1 = 0.3 \; l_2 = 0.4 \; l_3 = 0.3 \qquad (2)$$

We can update $p(l) = 1$ with our constrained Dempster's rule, to produce;

$$l_1 = 0.3 \; l_2 = 0.4 \; l_3 = 0.3 \quad \text{ie (2)}$$

But in view of the implicit constraints we must also update the refinements of m & n. We can do this by producing the equivalent of (2) in terms of m & n. Knowing that l_1 consists of m_1 & n_1 ...etc, we can derive from (2);

$$m_1 \,\&\, n_1 = 0.3 \;\; m_2 \,\&\, n_1 = 0.4 \;\; m_2 \,\&\, n_2 = 0.3 \qquad (3)$$

and separating with respect to the two refinement planes

$$m_1 = 0.3 \; m_2 = 0.4 \; m_2 = 0.3 \;\&\; n_1 = 0.3 \; n_1 = 0.4 \; n_2 = 0.3 \qquad (4)$$

which can update their respective states, with the constrained Dempster's rule, to produce;

$$m_1 = 0.3 \; m_2 = 0.7 \text{ and } n_1 = 0.7 \; n_2 = 0.3$$

Now, these results seem to reflect the sort of information that the evidence (2) suggests, and the method used

would seem to be robust and applicable in all possible situations. We must simply create new mass assignments from the evidence for all the refinement planes in which the evidence has constituent connections. We notice that the output does, in fact, allow the chance of m_1 & n_2 co-occuring - this is because there is not an equivalence between (3) and (4): (a & b) or (c & d) = (a or c) & (b or d) - but the probability assigned to this possibility is very low (0.09) and this sort of result represents, perhaps, the best way of enforcing that constituent constraint within a probabilistic lattice where contingent information can not be expressed. It also softens the hard and fast characteristics of the model in a way that, when we consider the knowledge-based derivation of its structure, is not that unwelcome. We stress, however, that this method is intuitive and has not, to our knowledge, received any formal investigation.

Finally, then, to the method of taking into account the interdependence between tests (C). This exists between tests that share evidential data. For Bayes or Dempster's rule to be valid, evidence, must be independent, and the best way to avoid the problem of test dependence would be just to have a single combined test which characterises a single combined feature space over all classes. There are two reasons why this is not done. The first, is that the model structures decisions into natural stages which facilitate the guide of phonetic knowledge and enable tests with non-intersecting domains to be trained independently. The second, is that estimating the large co-variance matrix required by a combined feature vector space would need a large and prohibitive amount of data. Assuming independence requires a far smaller amount of data, but might lead to problems in the reduction of information expressed by the statistical characterisation. A compromise, under consideration at the time of writing might be to produce a single figure value of tests' dependences, say;

$$\frac{\det[\text{covariance-matrix combined tests}]}{\det[\text{cov.mat. test1}] + \det[\text{cov.mat. test2}]}$$

which can then be used to reduce the effect of new evidence upon old in Dempster's combination rule.

This introduces a new problem, however. Namely; the combinatorial requirement of producing an interdependence value for every individual test with every possible combination of tests that could have preceded it. We can, however, consider a sensible subset of these values which enforce the minimum of constraints on the ordering of tests chosen by the scheduling system.

CONCLUSION

A method of evidential reasoning has been proposed for SYLK that compliments the structure of its syllable model. This involves a form of constrained Dempster's rule. EXTENSION to this, which satisfy the other constraints of the model have also been proposed, but await a serious investigation, and hopefully, justification.

REFERENCES

[1] P Cheeseman (1985), 'In defense of probability', Proc. IJCAI, Vol. 2, 1002-1009.
[2] A P Dempster (1968), 'A generalisation of Bayesian Inference', J. Royal Statistical Soc., Series B, 30, 205-247.
[3] P D Green et al (1990), '', in this volume
[4] E J Horvitz, D E Heckerman, C P Langlotz (1986), 'A framework for comparing alternative formalisms for plausible reasoning', Proc. Nat. Conf. Artificial Intelligence (AAAI-86) Vol. 1 (Science), 210-214.
[5] J T Nutter (1987), 'Uncertainty and probability', Proc. IJCAI, Vol. 1, 373-379.

[6] G Shafer (1976), *A Mathematical Theory of Evidence*, Princeton University Press.
[7] G Shafer (1981), 'Constructive probability', Synthese 48, 1-60.
[8] G Shafer and A Tversky (1985), 'Languages and designs for probability judgement', Cognitive Science, 9, 309-339.
[9] P M Williams (1976), 'Indeterminate probabilities', in: *Formal Methods in the Methodology of Empirical Sciences*, eds. M Przelecki, K Szanaiawski and R Wojzciki, Ossolineum and Reidel, 229-246.