Discarding irrelevant parameters in hidden Markov model based part-of-speech taggers

Eric Neufeld
Department of Computer Science
University of Saskatchewan
Saskatoon, SK, S7N 0W0
eric@spr.usask.ca

Abstract

A binary comparative definition of relevance, suggested by empirical results, gives a performance theory of relevance for hidden Markov models (HMMs) that makes it possible to reduce the total number of parameters in the model and while improving overall performance of the model in a specific application domain. Generalizations of this view of relevance are meaningful in many AI subareas.

Another view of this result is that there are at least two kinds of relevance. Knowledge of high quality is more relevant to a conclusion than low quality knowledge; specific knowledge is more relevant that general knowledge. This work argues that one can only be had at the expense of the other.

1 A relevance typology

When is knowledge "relevant"? When is knowledge "irrelevant"? There are at least two general notions in the AI community. There is a logicist notion of a premise being relevant to a conclusion, if, for example, it necessarily appears in a proof of the conclusion, as in relevance logic, or if its absence would lead to a differing, contradictory conclusion, as in the idea of "specificity" in nonmonotonic logic, as well as the distinction between contingent and background knowledge.

An opposite tack is taken by AI probabilists who focus on *irrelevance* rather than relevance. There are at least three probablistic notions of irrelevance: indifference, independence and randomness. *Indifference* is a relation on *probability distributions* (2) that assigns a background distribution, e.g., being indifferent to the outcome of a coin toss, each outcome is assigned probability 0.5. Given the probability a bird flies, indifference to a new variable *Red*, gives the probability a red bird flies with the same probability. *Independence* is a relation on *variables*. So, if a variable *Fly* is conditionally independent of *Bird* given *Red*, then the probability a red bird flies is the same as the probability a bird flies. Finally, *randomness* (4) is a relation on *outcomes of variables* that tells us to choose the narrowest reference class for which we have adequate statistics. Given a red bird, and the only applicable reference class is *bird* (i.e., nothing is known about red birds) we attach to our red bird the probability a bird flies. Note that for this simple example, all three notions generate the same result for the probability a red bird flies. As worlds involve more variables, the differences between the formalisms become apparent; this is not discussed here.

However, a point here is that different notions of relevance are functions of different abstractions of the data: as distributions, as variables, as outcomes of variables. For the present purpose, a single notion of relevance is studied in the next section. Interestingly, this generates a paradox of relevance, inspired by the experimental results.

2 A paradox of relevance

Clearly, knowledge is not relevant in the abstract, but is relevant to the degree of belief in a conclusion. Thus relevance is at least a binary relation on propositions. For example, a ringing smoke alarm is relevant to the conclusion my house is burning, if, say, the former increases the degree of belief in the latter. Independence relations on Bayes' nets suggest a ternary relation of conditional relevance, where a ringing smoke alarm is irrelevant to the conclusion my house is burning given the observation I observe there is no fire, although the alarm is relevant given no other information. More complex relations are possible.

Such notions of relevance are useful given perfect or idealized prior knowledge about the world. At one extreme, everything is relevant to everything else; at the other, all propositions are independent. Somewhere in between, a minimal set of relevance relations may imply many others. But knowledge of varying precision gives rise to an apparent *paradox of relevance*. On one hand, *more precise* knowledge is *more relevant* than less precise knowledge. On the other, *more specific* knowledge is more relevant than general knowledge. But one can only be had at the expense of the other, since more specific knowledge is generally based on

smaller samples. The practical question is, at what point should a handful of exceptions cease to be dismissed as errors and become an exceptional subclass?

This suggests a *binary comparative relevance* relation that defines when one sentence is *more relevant* than another to inferring the probability of a proposition. Then, given several statements, which is the most relevant?

Kyburg (4) suggests we use the narrowest class for which adequate statistics are available, where a statistic is adequate when its interval value does not completely contain the interval associated with a more general class, or that the difference between the two interval values cannot be explained away by error.

The empirical study below suggests this is also practical; the best performance results followed the realization that the most specific knowledge is not necessarily the most relevant.

3 HMM based part of speech tagging

A hidden Markov model trains on input-output pairs to learn parameters of an unobservable process that produces outputs. After training, the formalism is tested on new outputs. In the work described here, the model was trained on unrestricted natural language text tagged with parts of speech. After training, the model was tested on unseen text. The tag sequence attached to a word sequence by our tagger is that maximizing

$$\prod_{i=1}^{n} p(w_i|t_i)p(t_i|t_{i-1}) \qquad (1)$$

where w_i, t_i denote the ith word and tag type, respectively in an n-word sequence(5). Probabilities of the type $p(w_i|t_i)$ are called *lexical probabilities*; probabilities of the type $p(t_i|t_{i-1})$ are called *contextual probabilities*.

Neither the set of possible lexical probabilities nor the set of possible contextual probabilties is bounded by the training data and as a consequence a major problem is handling *unseen* words and tag sequences in the test data. The problem of unseen tag sequences is handled by the Good-Turing method; not discussed here. The problem of unseen words is significant. We found a 900000 token training set contained only 45000 unique words, half of which occurred exactly once. A disjoint 100000 token test set contained about 4000 words that had not been seen in the training set. Although some were specialized jargon, many (e.g., *outjumped*, *galaxy*) simply reflect the vastness of human experience.

In previous work, one technique for handling unseen words is using statistics collected from the training corpus whenever possible and revert to a secondary strategy based on other word features for unseen words. In (3) capitalization is used to identify nouns. In (6), statistics on a small predefined word ending set (e.g., *-ology*, *-ation*) are used. In (1), statistics on all 2-, 3- and 4-letter endings are used. A surprising result was that a tagger based on 3-letter endings alone outperformed taggers using 2- and 4- letter endings.

-ne -of -he -at -es -of -he -ic -ch -as -en -in -he -in -of -ch -ng, -he -ss -of -ng -ts -of -ch -to -ed -xt. -is -is -ly -ed -as -a -st -ep -ds -re -lt -ks -ch -as -ng, -ch -es -ic -ge.

(a) Sample text, all words truncated to two letters.

One of the great -es of the -ic approach has been in the -in of -ch -ng, the process of -ng parts of speech to -ed -xt. This is generally considered as a first -ep towards more difficult -ks such as -ng, which -es -ic knowledge.

(b) Same text, 1000 most frequent words added.

One of the great -es of the -ic approach has been in the -in of -ch -ng, the process of -ng parts of speech to -ed text. This is generally considered as a first step towards more difficult -ks such as -ng, which -es -ic knowledge.

(c) Same text, 3000 most frequent words added.

Figure 1: A little knowledge helps a lot.

If 3-letter ending knowledge outperforms 2-letter or 4-letter ending knowledge, 3-letter ending knowledge might outperform whole-word knowledge. This suggests the following novel approach: rather than tag words as a primary strategy and revert to an ending-based strategy for unseen words, tag endings as a primary strategy and revert to a word-based strategy for frequent words, for which the best possible statistics are available. Figure 1 gives the idea.

A strictly ending-based strategy doesn't outperform a word-based strategy, although performance is good. Using 1-letter endings results in a 70% success rate; 4-letter endings give about a 90 − −95% success rate and performance tends to improve with the ending length.

An unexpected result: performance improved, surpassing the word-based tagger by a significant 0.5%, and then *declined* as statistics based on small sample sizes were added. Figure 2 gives results, where entries `A`, `B`, `C`, `D` indicate different *effective tag list* strategies, (not discussed here), L is word-ending length, and N is number of whole words added back.

4 Conclusions

Here relevance is seen as trading quantity of knowledge against quality of knowledge. There are several ways of looking at this, e.g., as a way of providing a compact but efficient representations of domain structure, i.e., with the fewest parameters, as in neural nets. Or, as in nonmonotonic logic, we don't need to remember that

Exp.	N (×1000)					
	0	1	5	20	30	45
$L = 2$						
A	89.9	96.0	97.0	97.2	97.2	96.1
B	89.9	96.1	97.1	97.4	97.5	97.4
C	83.9	96.0	97.0	93.6	93.7	93.7
D	83.9	96.1	97.1	97.4	97.5	97.4
$L = 3$						
A	94.4	96.5	97.0	97.0	96.9	96.1
B	94.4	96.6	97.1	97.0	97.2	96.6
C	91.2	96.5	97.0	94.9	94.8	94.1
D	91.2	96.6	97.1	97.2	97.2	96.5

Figure 2: Tokens Correctly Tagged

each *kind* of bird flies, except penguins; we remember that birds fly but penguins don't. Nonmonotonic logic in this sense is efficient in that by focussing on what is relevant, it minimizes the number of parameters, although older treatments miss the issue of quality of knowledge. This idea of relevance also seems closely related to *overfitting* in machine learning — treating observational data as perfect information sacrifices flexibility. Yet another view: this relevance can be seen as aiding in deciding when one has acquired enough evidence to make a decision.

Another view is that we consider that there are two kinds of relevance — *ceteris paribus*, high quality knowledge is more relevant than low quality; similarly, more representative knowledge is more relevant than less. But one can be had only at the expense of the other.

References

[1] Greg Adams and Eric Neufeld. Automated word-class tagging of unseen words in text. In *Proceedings of the Sixth ISAI*, Mexico, 1993, pages 390–397.

[2] Peter Cheeseman. A method of computing generalized Bayesian probability values for expert systems. In *Proceedings of IJCAI-83*, 1983, pages 198-202.

[3] Kenneth W. Church. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Glasgow, U.K., May 1989.

[4] Henry E. Kyburg, Jr. The reference class. *Philosophy of Science*, 50:374–397, 1983.

[5] Bernard Merialdo. Tagging text with a probabilistic model. In *Proceedings of the IBM Natural Language ITL*, pages 161–172, Paris, 1990.

[6] Marie Meteer, Richard Schwartz, and Ralph Weischedel. POST: Using probabilities in language processing. In *IJCAI-91*, pages 960–965, Sydney, Australia, August 1991.

[7] Judea Pearl. 1988. Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann, California.