Tuning semantic consistency of active medical diagnosis: a walk on the semantic simplex

Albert Buchard, Adam Baker, Konstantinos Gourgoulias, Alexandre Navarro, Yura Perov, Max Zwiessele, and Saurabh Johri

AI Research Lab, Babylon Health

(Topic 6) Background. An increasing number of health-bots, able to perform diagnosis from a series of questions, are now available to patients. Active Medical Diagnosis describes the iterative process of hypothesis generation and hypothesis testing through which a model-based bot discovers the underlying cause of a patient's altered state. It is an online process which requires planning of a sequence of questions or tests, and dynamic re-evaluation as new evidence is gathered. Information Gain (IG) is a classic measure to rank the best tests to conduct during Active Diagnosis [7, 8]. As compared to rule-based [6, 11], or reinforcement learning algorithms [1, 9, 4, 2], expert systems based on greedily maximizing the IG [3, 10] produce question sequences close to optimal [3] without expert-crafted policy and without loss of interpretability. However, such an algorithm relies solely on the probabilistic properties of the underlying model and is blind to semantic knowledge. As a result, in patient-facing tools for self-diagnosis, this method may produce sequences characterized by seemingly abrupt changes of context [8], producing poor user experience and affecting the patient's trust in the system. Using a Bayesian Network (BN) crafted for medical diagnosis from case-history [5], we propose a method for embedding medical evidence over semantic simplices and compute a cost associated with changing contexts. Used in conjunction with greedy maximization, our approach allows clinicians and experts to tune the semantic consistency of question sequences without impacting the computational cost of the algorithm.

Method. Consider a 3-layer BN composed of risk-factors, diseases, and symptoms. We will denote the set of symptom and risk-factor nodes as the evidence set $\mathcal{S}$, through which evidence gets introduced into the model. Each node $E_i \in \mathcal{S}$ is associated with a unique question, and mapped by experts to specific semantic properties as well as to embeddings $s^{\mathrm{K}}_{E_i}$ over several semantic simplices $K \in \mathcal{K}$. Indeed, given a set Q of n independent properties, $Q = \{q_1, ..., q_n\}$, any subset $G \subseteq Q$ can naturally be mapped to a unique point s^G in a n-simplex, such that each coordinate satisfies

$$s^G_i = \begin{cases} \frac{1}{|G|}, & \text{if } q_i \in G, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

We define four such simplices from four medical taxonomies: coarse anatomical region (e.g., lower limb, $n = 19$), specific anatomical region (e.g., ankle, $n = 32$), pathogeny (e.g., infection, $n = 12$), and clinical system (e.g., Cardiology, $n = 15$). An isotropic energy landscape is then defined over each simplex K such that the barycenter, the point of greatest semantic ambiguity, has highest energy. Another key contribution of this work is the definition of an efficient cost function $C^{\mathrm{K}} : E_i, s^{\mathrm{K}}_t \to (-1, +1)$ over each semantic simplex K. The cost scales with the energy necessary to move to $s^{\mathrm{K}}_{E_i}$ given the current state s^{K}_t, defined as the average embedding of recent questions in K, and such that a trajectory towards the barycenter of the simplices incurs a positive cost (Figure 1). The greedy rule then picks the next evidence E^*,

$$E^* = \underset{E_i \in \mathcal{S} \setminus \mathcal{E}_{t-1}}{\operatorname{argmax}} \; IG(D|\mathcal{E}_{t-1}, E_i) - \sum_{K \in \mathcal{K}} w_{\mathrm{K}} C^K(E_i, s^K_t). \tag{2}$$

The IG is defined as the expected reduction of conditional entropy of a set of query diseases $D = \{d_0, \ldots, d_p\}$ after observing a new piece of evidence E_i. To account for the current state of

the diagnostic process, we also define a current evidence set attribution $\mathcal{E}_{t-1} = \{E_0 = e_0, \ldots, E_m = e_m\}, E_i \notin \mathcal{E}_{t-1}$. The hyperparameters w_{K} weight the costs over the different simplices, the results presented here used a non-zero weight for the Specific Anatomical simplex only. In order to evaluate the proposed algorithm A^* we compared it to a baseline algorithm A^0 relying solely on information gain without the cost. To produce the question sequences we ran both algorithms over a dataset of N=900 patient presentations, curated by clinicians. We judged the quality of the algorithm at improving semantic consistency by computing the *relative change in mean euclidean distance against an optimal sequence of questions produced by an oracle*. We also compared the quality of the algorithms over measures of efficiency at discovering clinical evidence, as well as diagnostic accuracy.

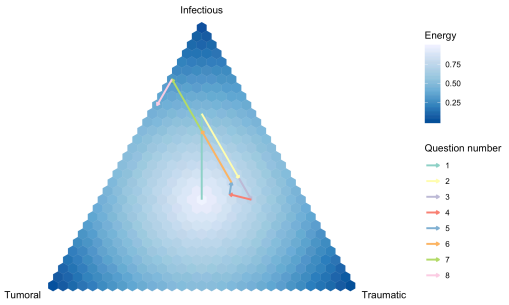

Figure 1: Sample trajectory of a question flow across a semantic 3-simplex with at each vertex a different pathogeny. The isotropic energy landscape is presented in color overlay.

	Classic VOI (A) N=900	Penalized VOI (A*) N=900	p adjusted (Bonferroni)
Efficiency	0.10 (0.08)	0.12 (0.08)	<0.001
Partial Efficiency	0.30 (0.21)	0.30 (0.19)	1.0
Accuracy	0.61 (0.49)	0.62 (0.49)	1.0
Relative Change From Optimal			
Coarse Anatomical	0.34 (0.49)	0.18 (0.45)	<0.001
Specific Anatomical	0.30 (0.75)	0.07 (0.67)	<0.001
Pathogeny	0.46 (0.59)	0.34 (0.48)	<0.001
Clinical System	0.26 (0.60)	0.25 (0.59)	1.0

Table 1: Evaluated metrics comparing A^* to the baseline A.

Results. A^* significantly improved semantic consistency over the weighted simplex (Specific Anatomy) but also over the non-weighted simplices: Coarse Anatomy and Pathogeny (Figure 2). The efficiency at discovering clinical evidence was also slightly but significantly improved, while partial efficiency, the ratio of question related to the true underlying disease, did not differ significantly. The accuracy did not differ significantly between the two approaches (Table 1).

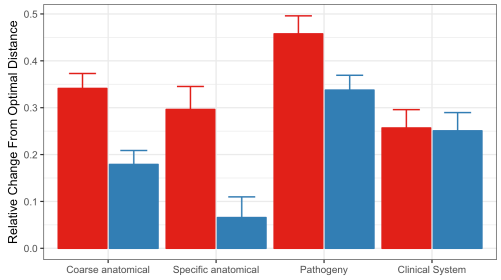

Figure 2: A^* significantly reduces the relative change in travelled distance from optimal over the Coarse Anatomy, Specific Anatomy, and Pathogeny simplex.

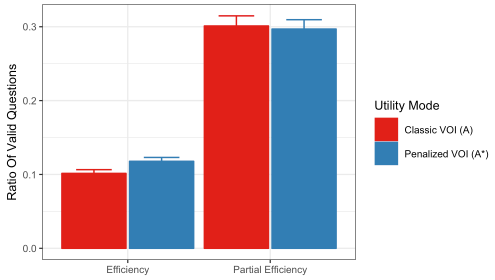

Figure 3: Efficiency, the ratio of questions asked relating to clinical evidence in the presentation, is significantly improved by A^*. Partial efficiency on the other hand did not differ significantly.

Implication for improving value of care. Along with improving the accuracy of expert systems for automated diagnosis, striving for improving the quality of the interaction with patients is essential to build trust in those systems. By introducing a cost incurred by changes of semantic contexts, our approach is able to improve the semantic consistency of question sequences produced

by greedy algorithms, without significant loss of accuracy or efficiency. This method is computationally efficient, linear in the number of potential questions, and can easily be interpreted and tuned by clinicians. In addition, it can be readily adapted to other types of decision algorithms, to non-myopic setting, to other semantic taxonomies, and can produce more complex behavior by introducing a dynamic energy landscape.

1. References

[1] Araya, M., Buffet, O., and Thomas, V. (2013). Active Diagnosis Through Information-Lookahead Planning. In *Journées Francophones de Planification, Décision et Apprentissage pour la conduite de systèmes*.

[2] Kao, H.-C., Tang, K.-F., and Chang, E. Y. (2018). Context-Aware Symptom Checking for Disease Diagnosis Using Hierarchical Reinforcement Learning. *Aaai*.

[3] Krause, A. and Guestrin, C. (2005). Optimal nonmyopic value of information in graphical models - Efficient algorithms and theoretical limits. In *IJCAI International Joint Conference on Artificial Intelligence*.

[4] Krishnamurthy, V. (2002). Algorithms for optimal scheduling and management of Hidden Markov model sensors. *IEEE Transactions on Signal Processing*, 50(6):1382–1397.

[5] Lucas, P. J., Van Der Gaag, L. C., and Abu-Hanna, A. (2004). Bayesian networks in biomedicine and health-care. *Artificial Intelligence in Medicine*, 30(3):201–214.

[6] Miller, R. A., Pople, H. E., and Myers, J. D. (1982). Internist-I , an Experimental Computer-Based Diagnostic Consultant for General Internal Medicine. *New England Journal of Medicine*, 307(8):468–476.

[7] Parsons, J. and Bao, L. (2018). The Value of Information in Retrospect.

[8] Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. Elsevier.

[9] Pellegrini, J. and Wainer, J. (2003). On the use of POMDPs to model diagnosis and treatment of diseases. *IV Encontro Nacional de Inteligência Artificial*.

[10] Rish, I., Brodie, M., Ma, S., Odintsova, N., Beygelzimer, A., Grabarnik, G., and Hernandez, K. (2005). Adaptive Diagnosis in Distributed Systems. *IEEE Transactions on Neural Networks*, 16(5):1088–1109.

[11] Shortliffe, E. H. (1977). Mycin: A Knowledge-Based Computer Program Applied to Infectious Diseases. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*.