Πολυτεχνείο Κρήτης

Σχολή

Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

ΣΥΣΤΗΜΑΤΑ ΣΤΑΤΙΣΤΙΚΗΣ ΔΙΑΧΕΙΡΙΣΗΣ ΔΙΑΛΟΓΟΥ

Επισκόπηση της Τεχνολογίας

Διπλωματική Εργασία

Κωνσταντίνος Πουλάκης

Εξεταστική Επιτροπή

Καθ. Μιχαήλ Γ. Λαγουδάκης

Αν. Καθ. Γεώργιος Χαλκιαδάκης

Δρ. Βασίλειος Διακολουκάς

Χανιά, Ιούλιος 2022

Technical University of Crete

School of Electrical and Computer Engineering

STATISTICAL DIALOGUE MANAGEMENT SYSTEMS

A survey of the technology

Diploma Thesis

Konstantinos Poulakis

Examination Committee

Prof. Michail G. Lagoudakis

Assoc. Prof. Georgios Chalkiadakis

Dr. Vasileios Diakoloukas

Chania, July 2022

ΠΕΡΙΛΗΨΗ

Τα σύγχρονα στατιστικά συστήματα διαχείρισης διαλόγου (Statistical Dialogue Managers - SDM) έχουν σημειώσει σημαντική πρόοδο, προσφέροντας εύρωστη και αποδοτική αλληλεπίδραση ανθρώπου-μηχανής. Η πρόοδος αυτή βασίζεται τόσο στο μεγάλο πλήθος δεδομένων, όσο και στην ανάπτυξη καινοτόμων αλγορίθμων βασισμένων σε τεχνικές ενισχυτικής μάθησης. Η εκτεταμένη χρήση αυτών των συστημάτων σε πραγματικά συστήματα φωνητικού διαλόγου (Spoken Dialogue Systems – SDS) μπορεί να μειώσει το κόστος ανάπτυξης και συντήρησης και να αυξήσει την ανοχή των συστημάτων στην αβεβαιότητα που υπάρχει τόσο λόγω περιβάλλοντος, όσο και λόγω σφαλμάτων των υποσυστημάτων που χρησιμοποιούνται όπως του συστήματος αναγνώρισης ομιλίας (Automatic Speech Recognition – ASR) ή του συστήματος κατανόησης φυσικής γλώσσας (Natural Language Understanding – NLU). Ένα τέτοιο αβέβαιο περιβάλλον, στο οποίο η κάθε απόφαση για την κατάσταση διαλόγου γίνεται σειριακά με άμεση εξάρτηση από τις προηγούμενες αποφάσεις, ένα από τα καταλληλότερα μοντέλα που χρησιμοποιηθεί βασίζεται σε μερικώς παρατηρήσιμες διαδικασίες απόφασης Markov (Partially Observable Markov Decision Process – POMDP). Στην πράξη, το μεγάλο πλήθος των καταστάσεων και των ενεργειών στον διάλογο, αλλά και η διάσταση των παρατηρήσεων, κάνει υπολογιστικά αδύνατη την βελτιστοποίηση του μοντέλου. Ως εκ τούτου, η πρακτική εφαρμογή συστημάτων που βασίζονται σε POMDP απαιτεί την ανάπτυξη αποτελεσματικών αλγορίθμων και προσεγγίσεων. Στην παρούσα διπλωματική εργασία επιχειρούμε μια αναλυτική επισκόπηση των μεθόδων και τεχνικών που έχουν αναπτυχθεί για την δημιουργία SDM. Αρχικά, επικεντρωνόμαστε στα συστήματα που βασίζονται σε POMDPs και εξετάζουμε διαφορετικές μεθόδους αναπαράστασης του χώρου των καταστάσεων του διαλόγου που έχουν ως στόχο την μείωση του υπολογιστικού κόστους και την βελτίωση των αποτελεσμάτων. Ακολούθως, γίνεται σύγκριση διαφορετικών μεθόδων μάθησης, τόσο γραμμικών όσο και μη γραμμικών, βασισμένων σε βαθιά νευρωνικά δίκτυα. Τα αποτελέσματα από μια σειρά πειραμάτων που έχουν διεξαχθεί στο περιβάλλον PyDial με χρήση τεχνητών δεδομένων από προσομοιωτή δείχνουν ότι η τεχνολογία είναι πολλά υποσχόμενη.

ABSTRACT

Modern Statistical Dialogue Managers (SDMs) have made significant strides in providing robust and efficient human-machine interaction. This progress is based both on the large amount of data and on the development of innovative algorithms based on augmented learning technique. Extensive use of these systems in real-world voice systems (Spoken Dialogue Systems (SDS)) can reduce development and maintenance costs and increase systems tolerance for uncertainty due to both environmental and subsystem errors used, such as Automatic Speech Recognition (ASR) or Natural Language Understanding (NLU). In such an uncertain environment in which each decision on the dialogue situation is made serially with direct dependence on previous decisions, one of the most appropriate models used is based on Partially Observable Markov Decision Process (POMDP). In practice, the large number of situations and actions in the dialogue, as well as the dimension of the observations, make it computationally impossible to optimize the model. Therefore, the practical implementation of POMDP-based systems requires the development of effective algorithms and approaches. In this diploma thesis, we attempt a detailed overview of the methods and techniques that have been developed to create SDM. We first focus on POMDPs-based systems and look at different methods of representing the dialog state space in order to reduce computational costs and improve results. Then, a comparison is made between different learning methods, both linear and non-linear based on deep neural networks. The results from a series of experiments conducted in the PyDial environment using artificial data from a simulator show that the technology is very promising.

Περιεχόμενα

ΠΙΝΑΚΑΣ ΕΙΚΟΝΩΝ

ΑΚΡΩΝΥΜΙΑ

Spoken Dialogue Systems (SDS)

Automatic Speech Recognition (ASR)

Natural Language Understanding (NLU)

Dialogue Manager (DM)

Natural Language Generator (NLG)

Text-to-Speech (TTS)

Interactive voice response (IVR)

Spoken Dialogue Managers (SDMs)

1 ΕΙΣΑΓΩΓΗ

Τα συστήματα διαλόγου είναι συστήματα που επιτρέπουν τη φυσική επικοινωνία μεταξύ ανθρώπων και μηχανών. Αυτή η επικοινωνία μπορεί να γίνει χρησιμοποιώντας κείμενο, φωνή, χειρονομίες ή άλλα φυσικά μέσα ανθρώπινης επικοινωνίας. Όταν η επικοινωνία ολοκληρώνεται χρησιμοποιώντας προφορική γλώσσα, τα συστήματα ονομάζονται Συστήματα Φωνητικού Διαλόγου (SDS). Τα SDS είναι σημαντικά, αφού για τους ανθρώπους η φωνή είναι το πιο φυσικό, γρήγορο και αποτελεσματικό μέσο επικοινωνίας. Ως εκ τούτου, η ερευνητική κοινότητα και οι επιχειρήσεις επενδύουν στην ανάπτυξη και την χρήση σε πραγματικά σενάρια των SDS για να πετύχουν φυσική αλληλεπίδραση ανθρώπου - μηχανής. Για παράδειγμα, μία από τις πρώτες εφαρμογές του SDS, είναι σε συστήματα που βελτιώνουν την προσβασιμότητα και την επικοινωνία των χρηστών με ειδικές ανάγκες ή των χρηστών που αντιμετωπίζουν δυσκολίες κινητικότητας. Ωστόσο, όλοι οι άνθρωποι μπορούν να επωφεληθούν από τέτοια συστήματα. Για παράδειγμα, οι χρήστες μπορούν να χρησιμοποιήσουν το SDS για να λάβουν πληροφορίες από τον Ιστό χρησιμοποιώντας προφορικά ερωτήματα, να ενημερωθούν ή να πραγματοποιήσουν συναλλαγές μέσω τηλεφώνου ή διαδικτύου, να στείλουν εντολές για τον έλεγχο έξυπνων συσκευών, να χρησιμοποιήσουν προσωπικούς βοηθούς, να υπαγορεύσουν μηνύματα ή κείμενο , για να χρησιμοποιήσετε το σύστημα πλοήγησης αυτοκινήτου και πολλές ακόμη εφαρμογές καθημερινής ζωής.

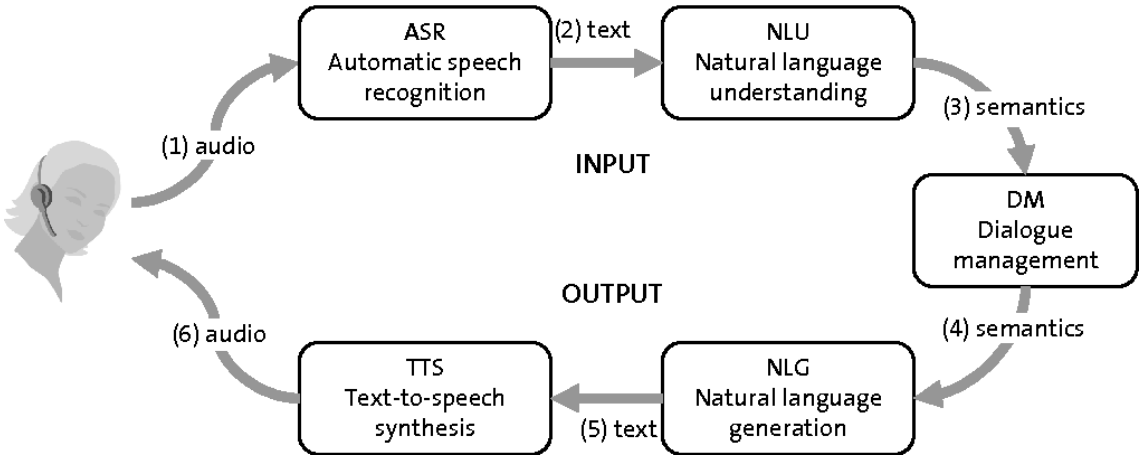
Η ανάπτυξη ενός SDS δεν είναι εύκολη υπόθεση. Απαιτεί τη αναπαράσταση της ανθρώπινης ομιλίας σε κάτι που μπορεί να γίνει κατανοητό από έναν υπολογιστή.

Ωστόσο, ένας υπολογιστής μπορεί να κατανοήσει μόνο μια σειρά από δυαδικά σήματα, ενώ ένας άνθρωπος έχει έναν περίπλοκο και διφορούμενο τρόπο να εκφραστεί μέσω της φωνής του. Συγκεκριμένα, υπάρχουν διαφορετικές προφορές, διάλεκτοι και γλώσσες. Το ίδιο άτομο εκφράζει πιθανότατα την ίδια σκέψη χρησιμοποιώντας διαφορετικές λέξεις ή τονισμό. Και η απόκριση του συστήματος θα πρέπει επίσης να λαμβάνει υπόψη το ιστορικό διαλόγου για να αποσαφηνίσει σωστά τις έννοιες. Αυτό που φαίνεται να είναι εύκολο για έναν άνθρωπο είναι ένα πολύ περίπλοκο πρόβλημα για έναν υπολογιστή.

Η διαδικασία της αλληλεπίδρασης ανθρώπου-μηχανής μέσω της φωνής απαιτεί τη συμμετοχή πολλών υποσυστημάτων που πρέπει να αλληλεπιδρούν απρόσκοπτα και να παρέχουν ακριβές αποτέλεσμα. Αυτά τα υποσυστήματα περιλαμβάνουν:

1. Το **Automatic Speech Recognition (ASR)** που μετατρέπει το σήμα ομιλίας στην αντίστοιχη σειρά λέξεων.
2. Το **Natural Language Understanding (NLU**) που εξάγει τη σημασιολογία από μια δεδομένη πρόταση με τη μορφή σημασιολογικών ετικετών.
3. Τον **Διαχειριστή διαλόγου (DM)** που βρίσκεται στην καρδιά του SDS και είναι υπεύθυνος να παρέχει την απόκριση του συστήματος με τη μορφή ενεργειών συστήματος.
4. Το **Natural Language Generator (NLG)** που μετατρέπει τις ενέργειες του συστήματος σε προτάσεις που είναι κατανοητές από τον άνθρωπο
5. Το **Text-to-Speech (TTS)** που μετατρέπει τις προτάσεις σε φωνή.

Ο τρόπος αλληλεπίδρασης αυτών των υποσυστημάτων φαίνεται στην **Error! Reference source not found.**.



Εικόνα 1: Τα υποσυστήματα ενός συστήμα

Όλα αυτά τα υποσυστήματα βασίζονται στην πλειονότητά τους σε στατιστικά μοντέλα. Αυτό κατέστη δυνατό λόγω της εκθετικής αύξησης των διαθέσιμων δεδομένων, σε συνδυασμό με την ανάλογη αύξηση της επεξεργαστικής ισχύος. Οι στατιστικές μέθοδοι προσφέρουν ένα πιο ευέλικτο, εύρωστο και συνεχώς βελτιούμενο πλαίσιο αλληλεπίδρασης που καθιστά δυνατή τη μεγάλη βελτίωση του SDS. Μια άλλη σημαντική ανακάλυψη από την οποία επωφελήθηκε σε μεγάλο βαθμό το SDS ήταν η ανάπτυξη αλγορίθμων βαθιάς μάθησης που βασίζονται σε νευρωνικά δίκτυα.

Υπάρχουν δύο κύριες κατηγορίες συστημάτων διαλόγου:

- Τα **task-oriented systems**: Είναι συστήματα σχεδιασμένα να επικοινωνούν για την ολοκλήρωση μιας συγκεκριμένης εργασίας, σε έναν συγκεκριμένο τομέα. Για παράδειγμα, ένα σύστημα διαδραστικής φωνητικής απόκρισης (IVR) που παρέχει τραπεζικές συναλλαγές και πληροφορίες ή βοηθά τον χρήστη να κλείσει ένα ιατρικό ραντεβού. Άλλοι δημοφιλείς τομείς περιλαμβάνουν την κράτηση σε ένα εστιατόριο, ένα ξενοδοχείο ή μια πτήση, πληροφορίες που σχετίζονται με ένα προϊόν σε ένα ηλεκτρονικό κατάστημα, όπως φορητό υπολογιστή ή λήψη πληροφοριών από έναν προσωπικό βοηθό, όπως το Google's Assistant και το Siri της Apple. Τα συστήματα αυτού του τύπου είναι τα πιο δημοφιλή αφού έχουν ένα ευρύ φάσμα εφαρμογών. Ως εκ τούτου, θα επικεντρωθούμε σε αυτή την κατηγορία στο υπόλοιπο της παρούσας διατριβής.
- Τα **non-task-oriented systems**: Η κύρια χρήση των συστημάτων που δεν προσανατολίζονται στην εργασία είναι για ψυχαγωγικούς και ψυχολογικούς λόγους. Συνήθως, δημιουργούν συνομιλίες μεταξύ ανθρώπων και υπολογιστών χωρίς να χρειάζεται να ολοκληρώσουν μια συγκεκριμένη εργασία. Επιπλέον, δεν περιορίζονται σε συγκεκριμένο τομέα. Απλώς παρέχουν γενική ανθρώπινη απάντηση, τις περισσότερες φορές με τη μορφή κειμένου, που μοιάζει με αλληλεπίδραση με έναν πραγματικό άνθρωπο. Συχνά ονομάζονται chatbot, καθώς χρησιμοποιούνται κυρίως για συνομιλία με ανθρώπους. Αν και υπάρχει αυξανόμενο ενδιαφέρον για την ανάπτυξη των chatbot, δεν είναι τόσο ευρέως διαδεδομένα και δεν υπάρχουν ακόμα πολλές εφαρμογές στον πραγματικό κόσμο. Όμως η εξέλιξη είναι τόσο σημαντική ώστε πρόσφατα το σύστημα LaMDA της Google επέδειξε δυνατότητες βαθιάς σκέψης και κατανόησης, ακόμα και έκφρασης ανθρώπινων συναισθημάτων που θα μπορούσαν να ταυτίζονται με την σκέψη ενός παιδιού επτά ετών.

Σε αυτή την εργασία θα επικεντρωθούμε κυρίως σε συστήματα προσανατολισμένα στην εργασία και θα δώσουμε ιδιαίτερη έμφαση στην ποικιλία των μεθόδων και τεχνικών που χρησιμοποιούνται για την κατασκευή του διαχειριστή διαλόγου.

1.1 Στατιστικός Διαχειριστής Διαλόγου

Ο ρόλος του διαχειριστή διαλόγου (DM) είναι πολύ σημαντικός σε ένα SDS. Είναι ένα σύστημα που καλείται να χειριστεί την αβεβαιότητα του διαλόγου τόσο λόγω περιβάλλοντος όσο και λόγω των σφαλμάτων στην αναγνώρισης ομιλίας από το ASR και των ασαφειών στη σημασία που μπορεί να οδηγήσουν σε σφάλματα κατανόησης από το NLU. Η δυσκολία που έχει ο DM στην διαχείριση ενός τυπικού διαλόγου γίνεται ακόμα περισσότερο κατανοητή αν λάβουμε υπόψη μας το γεγονός ότι ακόμα και ο ίδιος χρήστης σε ένα τέτοιο σύστημα, έχει τυπικά μια δυναμική απόκριση με διαφοροποιημένο τρόπο, λεξιλόγιο και συμπεριφορά στο πλαίσιο του ίδιου διαλόγου.
Σε τέτοια αβέβαια περιβάλλοντα, το έργο της επιλογής της βέλτιστης ενέργειας (dialogue action) του συστήματος σε κάθε κατάσταση διαλόγου (dialogue state) είναι μια εξόχως πολύπλοκη και δύσκολη διαδικασία. Η χρήση χειροποίητων κανόνων δεν είναι πάντα αποτελεσματική, καθώς τέτοια συστήματα είναι πολύ εύθραυστα στην αβεβαιότητα του διαλόγου και απαιτούν συνεχή επίβλεψη και ενημέρωση από τους προγραμματιστές. Επιπλέον, για εφαρμογές που περιέχουν πολλές καταστάσεις διαλόγου και πιθανές ενέργειες είναι πολύ κοστοβόρο και δύσκολο ή σχεδόν αδύνατο να δημιουργηθούν τέτοιοι χειροποίητοι κανόνες διαλόγου. Ως εκ τούτου, τα στατιστικά DMs (SDM) που βασίζονται σε δεδομένα έχουν ξεκινήσει να διερευνώνται από την ερευνητική κοινότητα ως μια καλή εναλλακτική και πλέον χρησιμοποιούνται σε αρκετές σύγχρονες εφαρμογές. Σε αυτή την εργασία εξετάζουμε διάφορες προσεγγίσεις SDM και δίνουμε μια εικόνα της τεχνολογίας που χρησιμοποιείται.

1.2 Πλεονεκτήματα

Κατά την κατασκευή ενός εμπορικής χρήσης συστήματος, συνήθως ξεκινάμε με την αποτύπωση των λειτουργικών αναγκών και τις προδιαγραφές της επιθυμητής συμπεριφοράς διαχείρισης διαλόγου και ακολούθως κατασκευάζεται ένας σκελετός του σύστημος ώστε να ταιριάζει με αυτό. Μόλις αναπτυχθεί η βασική συμπεριφορά του συστήματος, το σύστημα δοκιμάζεται σε πραγματικές συνθήκες με περιορισμένο αριθμό

χρηστών και στη συνέχεια προσαρμόζεται χειροκίνητα με στόχο την κάλυψη κενών στον διάλογο και την βελτίωση της ικανοποίησης των χρηστών. Η διαδικασία επαναλαμβάνεται μερικές φορές ακόμα μέχρι να επιτευχθεί ένα κατάλληλο επίπεδο ικανοποίησης του χρήστη και μπορεί να συνεχιστεί περιοδικά και μετά την εμπορική της χρήση ώστε να ενσωματώνονται μικρότερες ή μεγαλύτερες βελτιώσεις που μπορεί να απαιτούνται.

Εναλλακτικά, εφόσον χρησιμοποιηθούν δεδομένα και τεχνικές κατευθυνόμενης μάθησης για την διαμόρφωση του διαλόγου μπορεί να προκύψει βέλτιστη συμπεριφορά στην διαχείριση του διαλόγου αυτοματοποιημένα, αντλώντας πληροφορία απευθείας από τα δεδομένα φτάνει να υπάρχει μεγάλο όγκος δεδομένων διαλόγου. Με την συλλογή και χρήση πρόσθετων δεδομένων ο διαχειριστής διαλόγου μπορεί να προσαρμόσει την λειτουργία του και να βελτιώνεται συνεχώς ενώ παράλληλα μπορεί ένας διάλογος να προσαρμόζεται και να επεκτείνεται και σε νέες παρεμφερείς εφαρμογές.

Ωστόσο, η δυναμική φύση του διαλόγου οδηγεί σε μεγάλες δυσκολίες κατά την εφαρμογή τεχνικών που βασίζονται σε δεδομένα. Οι τομείς διαλόγου είναι συνήθως τουλάχιστον εκθετικοί ως προς τον αριθμό των διακριτών περιπτώσεων που μπορούν να δημιουργήσουν. Ακόμη και ένα πολύ μεγάλο διάλογος θα αντιπροσώπευε μόνο ένα μικρό κλάσμα του συνόλου των δυνατών διαλόγων. Ακόμα κι αν η συμπεριφορά του συστήματος μπορεί να γίνει εύκολα γνωστή, θα περιοριζόταν στη μίμηση μιας μορφής συμπεριφοράς σε μια συγκεκριμένη στροφή. Δεν υπάρχει καμία εγγύηση ότι μια τέτοια συμπεριφορά θα οδηγήσει σε έναν επιτυχημένο διάλογο.

Μια εναλλακτική είναι η χρήση της ενισχυτικής μάθησης, όπου ο διάλογος μοντελοποιείται ως διαδικασία διαδοχικής απόφασης και η συμπεριφορά διαχείρισης του διαλόγου βελτιστοποιείται σε σχέση με ένα αντικειμενικό μέτρο της απόδοσης του διαλόγου. Σε αντίθεση με την προσέγγιση εποπτευόμενης μάθησης, όπου η συμπεριφορά του διαχειριστή διαλόγου περιορίζεται μόνο σε αυτήν που εμφανίζεται στο σώμα, ένας διαχειριστής διαλόγου που χρησιμοποιεί ενισχυτική μάθηση μπορεί να διερευνήσει κάθε πιθανή συμπεριφορά. Είναι επομένως σε θέση να επιλέξει μια στρατηγική που βελτιστοποιεί τη συνολική απόδοση όπως ορίζεται από την μεγαλύτερη προσέγγιση των στόχων του διαλόγου.

1.3 Στόχος διπλωματικής εργασίας

Σε αυτή τη διπλωματική εργασία διερευνούμε διαφορετικά συστήματα Στατιστικών Διαχειριστών Διαλόγου (SDM). Έχουμε καταλάβει ότι ένα SDS είναι ένας προχωρημένος τρόπος αναζήτησης για Βάσεις Δεδομένων όπου η επικοινωνία ανθρώπου-μηχανής γίνεται μέσο διαλόγου. Η κύρια ασχολία ενός DM είναι να βάζει τιμές σε μεταβλητές του συστήματος και να δίνει ερωτήσεις και απαντήσεις στον χρήστη. Αυτό βλέπει ο χρήστης από εξωτερικά του συστήματος αλλά η δουλειά ενός DM, και αυτό που αναλύουμε στην διατριβή αυτή, είναι πιο πολύπλοκη καθώς χρειάζεται να βελτίωση και να επιταχύνει το διάλογο που χρησιμοποιείται. Όμως κάθε διάλογος είναι διαφορετικός οπότε δεν χρησιμοποιούμε αλγορίθμους και ντετερμινιστικά σχέδια για κάτι που δεν είμαστε σίγουροι. Οπότε χρησιμοποιούμε τεχνικές και ιδέες προς γενικευμένους στόχους. Ένα τέτοιο παράδειγμα είναι η μαρκοβιανή συνάρτηση που μας βοηθάει να πάρουμε απόφαση για τον πιο αποδοτικό τρόπο συνέχειας διαλόγου. Αυτή η συνάρτηση είναι ένα εξιδανικευμένο μοντέλο ενός διαλόγου που θεωρεί δεδομένα αρκετά πράγματα που δεν γνωρίζουμε στην αρχή (ή ποτέ). Αλλά αυτή την βασική ιδέα χρησιμοποιούμε σαν δομικό στοιχείο στις περισσότερες τεχνικές που θα δούμε. Αυτή είναι μια από τις συναρτήσεις που εμβαθύνουμε παρακάτω.

2 ΙΣΤΟΡΙΚΟ ΚΑΙ ΤΕΧΝΙΚΕΣ

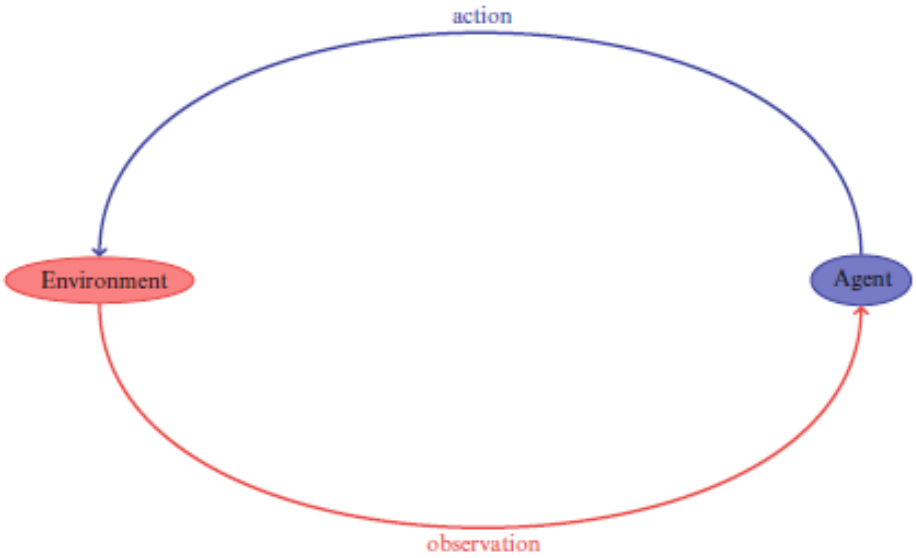
Σε αυτό το κεφάλαιο περιγράφουμε τις κύριες τεχνικές που χρησιμοποιούνται για την ανάπτυξη στατιστικών διαχειριστών διαλόγου (SDMs). Ο κύριος ρόλος ενός DM είναι να αποφασίσει ποια ενέργεια πρέπει να κάνει το σύστημα διαλόγου με βάση τις παρατηρήσεις και την υπάρχουσα κατάσταση στον διάλογο.

2.1 Ο διάλογος ως μια χρονικά μεταβαλλόμενη διαδικασία

Η υπόθεση σε όλες αυτές τις προσεγγίσεις είναι ότι η είσοδος στο DM που αντιστοιχεί άμεσα στην πρόθεση του χρήστη, είναι μια πληροφορία που μπορεί να είναι λανθασμένη. Ως εκ τούτου, ο στόχος του SDM είναι να αντιμετωπίσει την πιθανή αστοχία που μπορεί να εμφανιστεί στην είσοδό του και να διατηρήσει την ικανότητα ολόκληρου του συστήματος να εκπληρώσει τον γενικό του στόχο.

Για να παρέχει πληροφορίες στο χρήστη, ο διαχειριστής διαλόγου συνήθως χρησιμοποιεί μια τοπική βάση δεδομένων και/ή αναζητά δεδομένα στο Διαδίκτυο. Επιπλέον, λαμβάνει υπόψη πληροφορίες σχετικά με προηγούμενες καταστάσεις διαλόγου που ονομάζουμε το σημείο που βρίσκεται ο διάλογος κάποια χρονική στιγμή, οι οποίες διατηρούνται ως ιστορικό διαλόγου. Αυτές οι πληροφορίες είναι σημαντικές για να καθοδηγήσουν την απόφαση του διαχειριστή διαλόγου προς την ολοκλήρωση της αποστολής του. Για παράδειγμα, από τις πληροφορίες σε αυτήν την ενότητα, ο διαχειριστής διαλόγου μπορεί να παρατηρήσει ότι όλα τα δεδομένα σχετικά με μια κράτηση πτήσης αλλά την ημερομηνία αναχώρησης έχουν ήδη ληφθεί από τον χρήστη. Ως εκ τούτου, ο διαχειριστής διαλόγου μπορεί να αποφασίσει να ζητήσει από τον χρήστη τα δεδομένα που λείπουν.

Ο διάλογος σε αυτό το πλαίσιο μοντελοποιείται ως χρονοσειρά και το SDM είναι ένα σύστημα που λαμβάνει διαδοχικές αποφάσεις καθώς αλληλεπιδρά με το περιβάλλον. Η αλληλεπίδραση του συστήματος θεωρείται στοχαστική αφού το αποτέλεσμα της δράσης δεν είναι πλήρως γνωστό. Ένας τυπικός κύκλος αλληλεπίδρασης φαίνεται στο Σχήμα 2.1



Εικόνα 2: DM (Agent) and User (Environment) Interaction (Hamidreza Chinaei, 2016)

2.1.1 Διαδικασία Απόφασης Markov (MDP)

Το μαθηματικό πλαίσιο που χρησιμοποιείται συνήθως για τη μοντελοποίηση του διαλόγου είναι η Διαδικασία Απόφασης Markov (MDP). Το MDP χρησιμοποιείται πολύ στην τεχνητή νοημοσύνη (AI) για τη μοντελοποίηση και την επίλυση προβλημάτων σχεδιασμού αποφάσεων. Παρέχει ένα τρόπο απεικονίσεις της αλληλεπίδρασης ενός συστήματος με το περιβάλλον του, επιτρέποντάς μας να βρούμε τις καλύτερες στρατηγικές που μπορούν να καθοδηγήσουν αυτό το σύστημα μέσα στο περιβάλλον του. Στο παράδειγμα του SDS, το περιβάλλον του συστήματος αντιστοιχεί στον χρήστη. Άρα, ο στόχος ενός SDM που χρησιμοποιεί ένα MDP, είναι να βρει τις βέλτιστες στρατηγικές που μπορούν να καθοδηγήσουν έναν χρήστη στο περιβάλλον του.

Ας υποθέσουμε ότι έχουμε ένα ρομπότ που βρίσκεται σε μια περιοχή όπου μπορεί να μετακινηθεί. Το σύστημα βρίσκεται σε κατάσταση εκκίνησης (στην αρχή) και καθήκον του είναι να φτάσει σε μια τελική κατάσταση (στο τέλος).

Κάθε φορά που το σύστημα κάνει μια κίνηση προς τον στόχο του, αυτή η ενέργεια επηρεάζει το περιβάλλον και λαμβάνει μια ανταμοιβή (reward). Για να λειτουργήσει όσο το δυνατόν καλύτερα, το σύστημα θα πρέπει να σχεδιάσει ενέργειες που οδηγούν στην ολοκλήρωση της εργασίας και στην υψηλότερη μακροπρόθεσμη μέση ανταμοιβή. Στη βιβλιογραφία για την Τεχνητή Νοημοσύνη (AI), ένα τέτοιο σύστημα με συγκεκριμένο στόχο αναφέρεται ως πράκτορας (agent). Όταν η εργασία που πρέπει να ολοκληρώσει αυτός ο πράκτορας μοντελοποιείται ως Διαδικασία Απόφασης Markov, το πλαίσιο της τελευταίας ορίζεται ως το σύνολο των παραμέτρων $\{S, A, P, \gamma, R\}$, όπου::

- **Χώρος καταστάσεων (state space) S**: Το περιβάλλον του πράκτορα μοντελοποιείται από ένα σύνολο διακριτών καταστάσεων S. Γενικά το S μπορεί να είναι πεπερασμένο, μετρήσιμα άπειρο ή συνεχές. Σε αυτή την εργασία θεωρούμε το S ότι είναι πεπερασμένο με N,που συμβολίζει όλους τους φυσικούς αριθμούς, συνολικά καταστάσεις:$S = \{s_1, \dots, s_N\}$, με $N \in \mathbb{N}$..
- **Χώρος ενεργειών (action space) A**: Ένας πράκτορας που αλληλεπιδρά με το περιβάλλον του προσπαθεί να επηρεάσει το τελευταίο αναλαμβάνοντας ενέργειες από τον χώρο δράσης του A. Σε αυτή την εργασία, το A είναι πεπερασμένο:$A = \{\alpha_1, \dots, \alpha_M\}, M \in \mathbb{N}$.
- **Μοντέλο μετάβασης P**: Η συνάρτηση μετάβασης αποτυπώνει τη πιθανολογική φύση των επιδράσεων των ενεργειών. Έτσι, το P(s'|s,α) δίνει την πιθανότητα μετάβασης από την τρέχουσα κατάσταση s στην επόμενη s' δεδομένου ότι εκτελείται η ενέργεια α.
- **Συντελεστής έκπτωσης (discount factor) γ**: πραγματικός αριθμός μεταξύ 0 και 1.
- **Συνάρτηση ανταμοιβής (reward function) R**: Η συμπεριφορά του πράκτορα κωδικοποιείται στη συνάρτηση κόστους C(s,α) (έτσι ονομάζεται η αρνητική συνάρτηση ανταμοιβής), όπου χρησιμοποιείται όταν θεωρούμε την συνάρτηση R ως αρνητική. Η συνάρτηση κόστους αντιπροσωπεύει το άμεσο κόστος για την ανάληψη ενέργειας α όταν βρίσκεται σε κατάσταση s. Είναι ένας αλγόριθμος που χρησιμοποιεί το σύστημα για να μοντελοποιήσει την αποτελεσματικότητα μιας ενέργειας δεδομένης μιας συγκεκριμένης κατάστασης.

Στην τυπική υλοποίηση αυτού του πλαισίου, κάθε φορά που ο πράκτορας εκτελεί μια ενέργεια α όταν σε μια συγκεκριμένη κατάσταση s, το περιβάλλον αλλάζει στοχαστικά σύμφωνα με το μοντέλο μετάβασης P(s'| s,a) ως απόκριση σε αυτήν την ενέργεια. Καθώς ο πράκτορας μπορεί να χρησιμοποιήσει ένα περιορισμένο σύνολο ενεργειών A, χρησιμοποιεί τη συνάρτηση ανταμοιβής R(s,a) ως τρόπο να κρίνει την κατεύθυνση της συμπεριφοράς του και να αποφασίσει ποια ενέργεια πρέπει να γίνει. Με άλλα λόγια, η συνάρτηση ανταμοιβής αποτελείται από τους κανόνες που αποτελούν το περιβάλλον του πράκτορα. Και βοηθά τον πράκτορα να αξιολογήσει πόσο καλός ή κακός είναι ο αντίκτυπος των πράξεών του όταν βρίσκεται σε μια συγκεκριμένη κατάσταση.

Επομένως, ένα σημαντικό πλεονέκτημα είναι να υπάρχει ένα σύστημα που θα λαμβάνει υπόψη τον μελλοντικό αντίκτυπο των ενεργειών του. Σε μια τέτοια περίπτωση, η

συνάρτηση ανταμοιβής μπορεί να αναπαρασταθεί με διάφορους τρόπους. Για παράδειγμα, μπορεί να οριστεί ως το αναμενόμενο κόστος της δοκιμαστικής περιόδου:

$$Reward = \sum_{t=0}^{Tf} R(s_t, \alpha_t) \qquad (2.1)$$

το οποίο συνοψίζει όλες τις ανταμοιβές που βιώνει ο πράκτορας κατά τη διάρκεια μιας δοκιμαστικής συνεδρίας (μια διαδρομή στο χώρο κατάστασης που ξεκινά σε μια αρχική κατάσταση και τελειώνει σε μια τελική κατάσταση). Σε αυτήν την εξίσωση κατάσταση s_{Tf} , το Tf είναι το τελικό βήμα, στο οποίο επιτυγχάνεται μια τελική κατάσταση και το $R(s_t, a_t)$ είναι το κόστος που λαμβάνεται όταν η ενέργεια α_t πραγματοποιείται στην κατάσταση s_t. Μια παραλλαγή είναι το άπειρο χρονικό διάστημα, συνολικό μειωμένο κόστος:

$$Reward = \sum_{t=0}^{\infty} \gamma^t R(s_t, \alpha_t) \qquad (2.2)$$

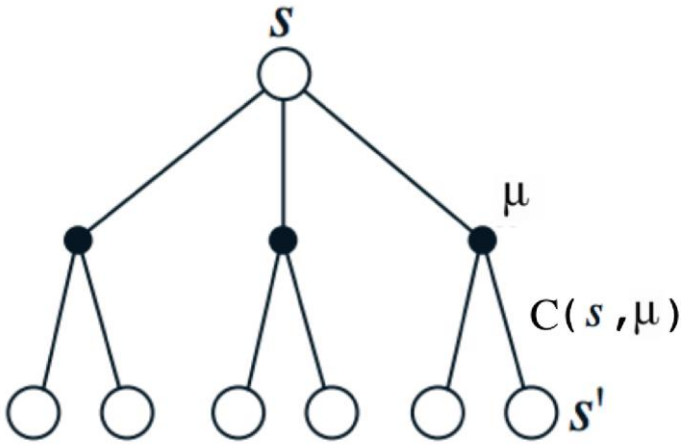
Τώρα που έχουμε καταλάβει τα βασικά στοιχεία του πλαισίου MDP και ο τρόπος που μπορεί να χρησιμοποιηθεί για τη μοντελοποίηση της αλληλεπίδρασης μεταξύ ενός πράκτορα και του περιβάλλοντος του έχει περιγραφεί, πρέπει να εστιάσουμε σε μια άλλη σημαντική έννοιά του: την πολιτική. Στον προγραμματισμό αποφάσεων, το πρόβλημα που αντιμετωπίζει ένας πράκτορας μπορεί να θεωρηθεί ως ποιες ενέργειες πρέπει να λάβει όταν βρίσκεται σε μια συγκεκριμένη κατάσταση. Η επιλογή που κάνει πράκτορας βασίζεται στην πολιτική (δηλαδή, στον κανόνα που ακολουθεί ο πράκτορας στην επιλογή ενεργειών, δεδομένης της κατάστασής του, μπορούμε να το θεωρήσουμε σαν μια λίστα ενεργειών για το καλύτερο $Reward$ που έχουμε βρει ως τώρα). Όπως καταλαβαίνουμε, η πολιτική δίνει τη λύση σε ένα MDP, δίνει μια πλήρη περιγραφή του ποια θα πρέπει να είναι η απόφαση του πράκτορα για κάθε κατάσταση s ϵ S. Έτσι, με δεδομένο ένα MDP, το καθήκον είναι να βρεθεί μια πολιτική που να μεγιστοποιεί το αναμενόμενο άθροισμα των ανταμοιβών που προκύπτουν από τις ενέργειες του πράκτορα.

Για την επίλυση του MDP, εισάγονται δύο ακόμη συναρτήσεις: η συνάρτηση τιμής V^{π}(s) και η συνάρτηση κόστους Q^{π} (s,α), για κάθε κατάσταση s και ενέργεια α.

Συνάρτηση τιμής: Ας υποθέσουμε ότι έχουμε ένα MDP, ένας χώρος καταστάσεων S, ένας χώρος δράσης A, ένα μοντέλο μετάβασης P (s' | s, α) και μια συνάρτηση ανταμοιβής R (s, α). Ας υποθέσουμε ότι έχουμε επίσης μια πολιτική π (π: S→A), τέτοια ώστε π (s) να είναι η ενέργεια α που πρέπει να κάνει ο πράκτορας στην κατάσταση s. Συγκεκριμένα, μια πολιτική π είναι μια αντιστοίχιση από κάθε κατάσταση s ϵ S στην ενέργεια α ϵ A. Ο Bellman ορίζει ότι η τιμή μιας κατάστασης s στην πολιτική π, που συμβολίζεται με V^{π}(s),

είναι το αναμενόμενο κόστος όταν ξεκινάει από την κατάσταση s και ενεργώντας σύμφωνα με την πολιτική π στη συνέχεια. Για τα MDP, μπορεί να διατυπωθεί ως εξής:

$$\boldsymbol{V^{\pi}(s) = \sum_{s' \epsilon S} P(s' \,|s, \pi(s))[R(s, \pi(s)\,) + V^{\pi}(s')]} \qquad \textbf{(2-1)}$$



Εικόνα 3: Backup diagram for V^{π}

Ένας τρόπος για να σκεφτείτε αυτή τη συνάρτηση, είναι να σκεφτείτε εκ των προτέρων από μια κατάσταση σε όλες τις πιθανές καταστάσεις των διαδόχων της, όπως απεικονίζεται στο Σχήμα 2.2. Σε αυτό το διάγραμμα, οι άδειοι κύκλοι αντιστοιχούν σε μια κατάσταση και οι γεμάτοι κύκλοι αντιστοιχούν σε ένα μικρό διάνυσμα κατάστασης-δράσης (για παράδειγμα, (s, α)). Ξεκινώντας από την κατάσταση s, ο πράκτορας θα μπορούσε να κάνει οποιαδήποτε ενέργεια από τον χώρο ενεργειών του. Κάθε μία από αυτές τις τρεις ενέργειες, από την αρχική κατάσταση, θα μπορούσε στη συνέχεια να συνδυαστεί με s και να αντιστοιχεί σε έναν συμπαγή κύκλο. Η ανάληψη δράσης σε μια συγκεκριμένη κατάσταση συνεπάγεται κόστος (ανταμοιβή). Έτσι, αν υποθέσουμε ότι όταν στην κατάσταση s ο πράκτορας ακολουθεί την πολιτική π και αναλαμβάνει δράση α, τότε προκύπτει ένα άμεσο και βέλτιστο κόστος και ο πράκτορας μετακινείται από την κατάσταση s σε πιθανές άλλες επόμενες καταστάσεις, όπως το s'.

Η εξίσωση (2-1) δίνει μέσο όρο για όλες τις πιθανότητες, σταθμίζοντας το καθένα με την πιθανότητα να συμβεί. Η εξίσωση δηλώνει ότι η τιμή της κατάστασης s ισούται με την αναμενόμενη τιμή του επόμενου βήματος s' που συσσωρεύτηκε στην πορεία, συν την άμεση ανταμοιβή (δηλαδή, R(s, α) στο (2-2)). Υπάρχει σχεδόν πάντα μία πολιτική π* που είναι καλύτερη ή ίση από όλες τις άλλες πολιτικές. Αυτή η πολιτική ονομάζεται βέλτιστη πολιτική. Η βέλτιστη συνάρτηση τιμής που βασίζεται σε αυτή τη βέλτιστη πολιτική δίνεται στην αρχή της βελτιστοποίησης του Bellman:

$$\boldsymbol{V^*(s) = \sum_{s' \epsilon S} P(s' \mid s, \alpha)[R(s, \alpha) + V^*(s')]} \qquad \textbf{(2-2)}$$

Η βέλτιστη τιμή του s, V* (s), είναι η αναμενόμενη ανταμοιβή όταν ξεκινάτε από την κατάσταση s και ενεργείτε σύμφωνα με τη βέλτιστη πολιτική π*.

Συνάρτηση Q: Τώρα, έστω η τιμή Q ενός ζεύγους κατάστασης-δράσης, που συμβολίζεται με Q^{π} (s, α), είναι το βέλτιστο κόστος για την ανάληψη δράσης α όταν βρίσκεται στην κατάσταση s και μετά την ενέργεια σύμφωνα με την πολιτική π. Βλέπουμε ότι τα Q^{π} (s,α) και V^{π} (s) μπορούν να οριστούν αναδρομικά το ένα ως προς το άλλο. Αυτό μπορεί να διαμορφωθεί ως εξής:

$$\boldsymbol{Q^{\pi}(s, \alpha) = \sum_{s' \epsilon S} P(s' \mid s, \alpha)[R(s, \alpha) + V^{\pi}(s')]} \qquad (2.5)$$

Με

$$\boldsymbol{Q^{\pi}\big(s, \pi(s)\big) = V^{\pi}(s)} \qquad (2.6)$$

Αυτή η εξίσωση δηλώνει ότι εάν ο πράκτορας βρίσκεται στην κατάσταση s, αναλάβει δράση α και μετακινηθεί στην κατάσταση s', επιφέρει την άμεση ανταμοιβή R(s,α), συν την αναμενόμενη ανταμοιβή V^{π} (s') που συσσωρεύεται κατά μήκος του τρόπος. Βλέπουμε ότι, μετά την εκτέλεση της ενέργειας α στην κατάσταση s, ο πράκτορας θα μπορούσε να μετακινηθεί από το s σε οποιαδήποτε πιθανή επόμενη κατάσταση, ενώ θα επιφορτώθει με το αντίστοιχο άμεσο κόστος. Ας υποθέσουμε ότι μετά την εκτέλεση της ενέργειας α ενώ βρίσκεται στην κατάσταση s, ο πράκτορας επιβαρύνεται με ένα κόστος, και στη συνέχεια φτάνει στην επόμενη κατάσταση s'. Ακολουθώντας τον ορισμό του Q^{π} (s,α), όταν ο πράκτορας φτάνει στην κατάσταση s΄, συνεχίζει ακολουθώντας την πολιτική π, και αναλαμβάνει δράση α΄, με π(s΄) = α΄. Έτσι, η βέλτιστη συνάρτηση Q^{π}: Q* (s, α) ορίζεται ως η αναμενόμενη ανταμοιβή που επιστρέφεται όταν η ενέργεια α εκτελείται στην κατάσταση s, και η βέλτιστη πολιτική π* ακολουθείται στη συνέχεια:

$$\boldsymbol{Q^*(s, \alpha) = \sum_{s' \epsilon S} P(s' \mid s, \alpha)[R(s, \alpha) + V^*(s')].} \qquad (2,7)$$

Επομένως, η συνάρτηση βέλτιστης τιμής μπορεί επίσης να εκφραστεί ως εξής:

$$\boldsymbol{V^*(s) = Q^*(s, \alpha)}\ . \qquad (2.8)$$

Η βέλτιστη πολιτική π* μπορεί στη συνέχεια να προκύψει από τη συνάρτηση βέλτιστης τιμής με:

$$\boldsymbol{\pi^* = arg[\sum_{s' \epsilon S} P(s' \mid s, \alpha)(R(s, \alpha) + V^*(s'))]} \qquad (2.9)$$

ή από τη βέλτιστη συνάρτηση Q:

$$\boldsymbol{\pi^* = arg Q^*(s, \alpha)}$$

2.1.2 Μερικώς Παρατηρήσιμη MDP (Partial Observable Markov Decision Process)

Αν και το MDP έχει αποδειχθεί ένα αποτελεσματικό σχήμα μοντελοποίησης σε πολλές προσεγγίσεις, βασίζεται στην υπόθεση ότι το περιβάλλον διαλόγου είναι πλήρως παρατηρήσιμη. Ωστόσο, αυτό δεν συμμορφώνεται με τις περισσότερες πραγματικές εφαρμογές, στις οποίες η αβεβαιότητα οδηγεί σε ένα μερικώς παρατηρήσιμη πλαίσιο διαλόγου. Ως εκ τούτου, τα Μερικώς Παρατηρήσιμα MDP (POMDP) έχουν επίσης διερευνηθεί για την αντιμετώπιση αυτού του περιορισμού. Το POMDP είναι ένα πιο γενικευμένο πλαίσιο σχεδιασμού σε συνθήκες αβεβαιότητας, όπου η βασική υπόθεση είναι ότι οι καταστάσεις είναι μόνο εν μέρει παρατηρήσιμες. Ένα POMDP αντιπροσωπεύεται ως πλειάδα $\{S, A, P, \gamma, R, O, \Omega, b_0\}$. Δηλαδή, ένα μοντέλο POMDP περιλαμβάνει ένα μοντέλο MDP και επίσης προσθέτει:

- **Χώρος παρατήρησης O**: Ένα σύνολο παρατήρησης όπου ο ϵ O. Το οποίο είναι μια παρατήρηση του συστήματος μετά την ερμηνεία της τρέχουσας ενέργειας α.
- **Πιθανότητα παρατήρησης Ω**: Πιθανότητα παρατήρησης όπου $\Omega(o', s', \alpha) = P(o' \mid \alpha, s')$.
- **Αρχική κατάσταση πεποίθησης b_0**: Ένα σύνολο πιθανοτήτων που δείχνει την πεποίθηση του πράκτορα για την κατάσταση s_0.

Για να βάλουμε ο αριθμητική τιμή στο b έχουμε μια συνάρτηση Εκτιμητή Κατάστασης $SE(b, \alpha, o')$

$$\boldsymbol{b'(s') = SE(b, \alpha, o') = Pr\ Pr\ (s' \mid b, \alpha, o') = \eta \Omega(\alpha, s', o') \sum_{s \in S} b(s) P(s, \alpha, s')}$$

όπου (η) είναι ο παράγοντας κανονικοποίησης, που ορίζεται ως:

$$\boldsymbol{\eta = \frac{1}{Pr(o' \mid b, a)}}$$

$$\boldsymbol{Pr\ Pr\ (b, a) = \sum_{s' \in S} \Omega(\alpha, s', o') \sum_{s \in S} b(s) P(s, \alpha, s')}$$

Η συνάρτηση ανταμοιβής μπορεί επίσης να οριστεί στις πεποιθήσεις:

$$\boldsymbol{R(b, \alpha) = \sum_{s \in S} b(s) R(s, \alpha)} \qquad (2.12)$$

Η πολιτική POMDP επιλέγει μια ενέργεια a για μια κατάσταση πεποίθησης b, δηλαδή α=π(b). Στο πλαίσιο POMDP ο στόχος είναι να βρεθεί μια βέλτιστη πολιτική π*, όπου για οποιαδήποτε πεποίθηση b, η π* καθορίζει μια ενέργεια α=π(b) που μεγιστοποιεί την αναμενόμενη τιμή των μελλοντικών ανταμοιβών ξεκινώντας από την πεποίθηση b0. Είναι επίσης μια εξίσωση Bellman για τα POMDP:

$$V^{\pi}(b) = [R\big(b, \pi(b)\big) + \gamma \sum_{o' \in O} Pr\big(o' \big| b, \pi(b)\big) V^{\pi}(b')]. \quad (2.13)$$

Και το βέλτιστο μοντέλο πεποίθησης-αξίας V* μπορεί να βρεθεί από το $V^*=V^{\pi}(b)$.

Παρατηρήστε ότι μπορούμε να δούμε ένα POMDP ως MDP, εάν το POMDP περιλαμβάνει ένα μοντέλο παρατήρησης και μια αρχική πεποίθηση. Αυτό φαίνεται στην Εξ. (2.11), ξεκινώντας με μια αρχική πεποίθηση, η επόμενη πεποίθηση θα είναι ντετερμινιστική καθώς το μοντέλο παρατήρησης είναι ντετερμινιστικό. Αυτό σημαίνει ότι ένα τέτοιο POMDP που γνωρίζει την τρέχουσα κατάστασή του με 100% πιθανότητα είναι παρόμοιο με τα MDP.

2.2 Αναπαράσταση χώρου κατάστασης

Μια αναπαράσταση χώρου κατάστασης είναι μια απλή μαθηματική αναπαράσταση ενός φυσικού συστήματος ως ένα σύνολο μεταβλητών εισόδου, εξόδου και κατάστασης που σχετίζονται με διαφορικές εξισώσεις πρώτης τάξης ή εξισώσεις διαφοράς. Οι μεταβλητές κατάστασης των οποίων οι τιμές εξελίσσονται με την πάροδο του χρόνου με τρόπο που εξαρτάται από τις τιμές που έχουν κάθε δεδομένη χρονική στιγμή και επίσης εξαρτάται από τις εξωτερικά εισαγόμενες τιμές των μεταβλητών εισόδου. Άρα οι μεταβλητές καταστάσεων επηρεάζουν όλο το σύστημα.

Κάθε κατάσταση διαλόγου μοντελοποιείται ως διάνυσμα κατάστασης, προκειμένου να χρησιμοποιηθεί από τη διαχείριση διαλόγου.

2.2.1 Διάνυσμα καταστάσεων ολικής πεποίθησης

Ένας ανιχνευτής πεποιθήσεων χρησιμοποιείται για την οικοδόμηση της πεποίθησης για το διάνυσμα κατάστασης.

Σε κάθε χρονικό βήμα, η μηχανή βρίσκεται σε μια συγκεκριμένη κατάσταση st, αλλά εφόσον η κατάσταση είναι μη παρατηρήσιμη, η μηχανή διατηρεί μια κατανομή σε όλες τις πιθανές καταστάσεις τη χρονική στιγμή t – που ονομάζεται κατάσταση πεποίθησης, $b(s_t)$. Τότε, η πιθανότητα η μηχανή να βρίσκεται στην κατάσταση s,s ϵ S, τη στιγμή t είναι $b(s_t = s)$. Επομένως, η κατάσταση πεποίθησης $b(s_t)$ παίρνει τιμές b ϵ B όπου B = [0, 1]|S| είναι ο χώρος των πεποιθήσεων. Η αρχική κατανομή στις καταστάσεις $b(s_0)$ δίνεται από το διάνυσμα $b^0 = [b(s_0= s^1),...,b(s_0= s^{(|S|)})]^T$. Η αρχική κατανομή b_0, μαζί με τις πιθανότητες παρατήρησης P(s',o'), τις πιθανότητες μετάβασης Ps=s'(a) και τις προσδοκίες ανταμοιβής $R(s_a)$, προσδιορίζει πλήρως το μοντέλο POMDP.

Όταν φθάνει μια νέα παρατήρηση, η νέα κατάσταση πεποίθησης b(s(t+1)) λαμβάνεται ως η κατανομή πιθανότητας σε όλες τις πιθανές καταστάσεις τη χρονική στιγμή t + 1, δεδομένης της παρατήρησης o_{t+1}, της ενέργειας που έγινε και της προηγούμενης κατάσταση πεποίθησης b(st). Με την προϋπόθεση ότι ο χώρος καταστάσεων είναι πεπερασμένος, η νέα κατάσταση πεποίθησης μπορεί να υπολογιστεί ως εξής:

$$\boldsymbol{b(s_{t+1} = s')}$$

$$\boldsymbol{= P(o_{t+1} = o', a_t = a, b(s_t) = b)}$$

$$\boldsymbol{= \frac{P(s_{t+1} = s', a_t = a, b(s_t) = b)P(a_t = a, b(s_t) = b)}{P(a_t = a, b(s_t) = b)}}$$

$$\boldsymbol{\propto P(s_{t+1} = s') \sum_{s \in S} P(a_t = a, b(s_t) = b, s_t = s) P(a_t = a, b(s_t) = b)}$$

$$\boldsymbol{= P_{s'o'} \sum_{s \in S} P^{a}_{ss'} b(s_t = s)},$$

για κάθε s'ϵ S, όπου η κατάσταση πεποίθησης τη χρονική στιγμή t, $b(s_t)$, δίνεται από το $b= [b(s_t= s^1),...,b(s_t= s^{(|S|)})]^T$, η ενέργεια a πραγματοποιείται τη στιγμή t και η παρατήρηση o' παρατηρείται τη στιγμή t + 1.

2.2.2 Κατάσταση αθροιστικής πεποίθησης

Ευτυχώς, υπάρχουν ορισμένοι αφαιρετικοί περιορισμοί που μπορούν να αξιοποιηθούν. Πρώτον, μόνο ένα σχετικά μικρό μέρος του χώρου πεποιθήσεων θα χρησιμοποιηθεί πραγματικά κατά τη διάρκεια οποιουδήποτε κανονικού διαλόγου, και δεύτερον, το εύρος των λογικών ενεργειών σε οποιοδήποτε συγκεκριμένο σημείο του χώρου πεποιθήσεων θα είναι συχνά περιορισμένο. Αυτό εισάγει την έννοια ενός συμπιεσμένου χώρου χαρακτηριστικών που ονομάζεται χώρος σύνοψης στον οποίο απλοποιούνται τόσο οι καταστάσεις όσο και οι ενέργειες προκειμένου να επιτραπεί η αναπαράσταση και η βελτιστοποίηση μίας πολιτικής με δυνατότητα μεταφοράς. Ο χώρος σύνοψης μπορεί επομένως να θεωρηθεί ως ένας υποχώρος του πλήρους κύριου χώρου όπου η παρακολούθηση πεποιθήσεων εκτελείται στον κύριο χώρο και η λήψη αποφάσεων και η βελτιστοποίηση πολιτικής πραγματοποιούνται σε χώρο σύνοψης. Η λειτουργία χρόνου εκτέλεσης ενός χώρου κύριας σύνοψης POMDP είναι επομένως η εξής. Μετά την ενημέρωση πεποιθήσεων, η κατάσταση πεποίθησης b στον κύριο χώρο αντιστοιχίζεται σε ένα διάνυσμα χαρακτηριστικών {b'} και σε ένα αντίστοιχο σύνολο υποψήφιων ενεργειών {a'}. Στη συνέχεια, η πολιτική χρησιμοποιείται για την επιλογή της καλύτερης

ενέργειας για την εκτέλεση b'→a' από το σύνολο των υποψήφιων ενεργειών και μια δεύτερη ευρετική χρησιμοποιείται για να αντιστοιχίσει το a' σε μια πλήρη ενέργεια a στον κύριο χώρο.

Το POMDP λειτουργεί ως εξής. Σε κάθε χρονικό βήμα, ο κόσμος βρίσκεται σε κάποια απαρατήρητη κατάσταση s_t. Εφόσον το s_t δεν είναι ακριβώς γνωστό, διατηρείται μια κατανομή σε πιθανές καταστάσεις που ονομάζεται κατάσταση πεποίθησης b_t όπου το $b_t(s_t)$ δείχνει την πιθανότητα να βρίσκεται σε μια συγκεκριμένη κατάσταση s_t. Με βάση το b_t, το μηχάνημα επιλέγει μια ενέργεια στο, λαμβάνει μια ανταμοιβή r_t και μεταβαίνει στην (μη παρατηρούμενη) κατάσταση s_{t+1}, όπου το s_{t+1} εξαρτάται μόνο από το s_t και το a_t. Στη συνέχεια, η μηχανή λαμβάνει μια παρατήρηση o_{t+1}, η οποία εξαρτάται από s_{t+1} και a_t.

Σε ένα πρακτικό SDS προσανατολισμένο στην εργασία, η κατάσταση πρέπει να κωδικοποιεί τρεις διαφορετικούς τύπους πληροφοριών: τον στόχο του χρήστη g_t, την πρόθεση της πιο πρόσφατης εκφοράς χρήστη u_t και το ιστορικό διαλόγου h_t. Ο στόχος περιλαμβάνει τις πληροφορίες που πρέπει να συλλεχθούν από τον χρήστη προκειμένου να εκπληρωθεί η εργασία, η πιο πρόσφατη έκφραση χρήστη αντιπροσωπεύει αυτό που πραγματικά ειπώθηκε σε αντίθεση με αυτό που αναγνωρίστηκε και το ιστορικό παρακολουθεί σχετικές πληροφορίες που σχετίζονται με προηγούμενες στροφές. Η παραγοντοποίηση της κατάστασης με αυτόν τον τρόπο είναι χρήσιμη επειδή μειώνει τις διαστάσεις του πίνακα μετάβασης κατάστασης και μειώνει τον αριθμό των εξαρτήσεων υπό όρους.

Υπάρχουν πιθανές παραλλαγές σε αυτή την παραγοντοποίηση. Για παράδειγμα, η επίδραση του χρήστη μπορεί επίσης να παραγοντοποιηθεί, αλλά οι περισσότερες τρέχουσες προσεγγίσεις ταιριάζουν σε αυτό το μοντέλο. Στις προσεγγίσεις N-best, η κατάσταση πεποίθησης προσεγγίζεται από μια λίστα με τις πιο πιθανές καταστάσεις με τις πιθανότητές τους. Αυτό σημαίνει ότι οι καταστάσεις διαλόγου που αντιστοιχούν στις πιο πιθανές ερμηνείες της πρόθεσης του χρήστη είναι καλά μοντελοποιημένες, με άλλες καταστάσεις να δίνουν χαμηλή πιθανότητα μάζας. Ένα παράδειγμα αυτής της προσέγγισης είναι ο αλγόριθμος Κρυφής Κατάστασης Πληροφοριών (HIS) το οποίο ομαδοποιεί παρόμοιους στόχους χρήστη σε κλάσεις ισοδυναμίας που ονομάζονται κατατμήσεις με την υπόθεση ότι όλοι οι στόχοι στο ίδιο διαμέρισμα είναι εξίσου πιθανοί.

Τα τμήματα είναι δομημένα σε δέντρο ώστε να απαιτούν την εκτίμηση των συνθηκών που χαρακτηρίζονται στη μεταφυσική του χώρου και είναι κατασκευασμένα χρησιμοποιώντας σύνολα τιμών θυρίδων από τη λίστα N-best λίστα των θεωριών

αναγνώρισης και την τελική απόδοση πλαισίου. Ο συνδυασμός ενός τμήματος, μιας πράξης πελάτη από τη λίστα N-best και του σχετικού ιστορικού ανταλλαγής σχηματίζει μια εικασία. Μια διασπορά πιθανοτήτων στις πιο πιθανές εικασίες διατηρείται εν μέσω της ανταλλαγής και αυτό αποτελεί τον χώρο πεποίθησης. Η παρατήρηση πεποίθησης σε εκείνο το σημείο απαιτεί όπως ήταν να ενημερωθούν οι πεποιθήσεις της υπόθεσης και δεδομένου ότι υπάρχουν σχετικά λίγες θεωρίες, αυτό θα μπορούσε αβίαστα να εξαντληθεί στον πραγματικό χρόνο. Η επιλογή για τη διατήρηση μιας N-best λίστας στόχων πελάτη είναι να υπολογιστεί ο στόχος του πελάτη σε έννοιες που μπορούν να συζητηθούν σχεδόν από το πλαίσιο

Αυτό δείχνει τους διαφορετικούς συμβιβασμούς μεταξύ της προσέγγισης N-best, η οποία μπορεί να μοντελοποιήσει όλες τις εξαρτήσεις, αλλά με μια ατελή κατανομή, και την προσέγγιση του factoring σε επίπεδο slot, η οποία μπορεί να χειριστεί μόνο έναν περιορισμένο αριθμό εξαρτήσεων, αλλά μπορεί να μοντελοποιήσει την πλήρη κατανομή.

2.2.3 Τμηματικά Ανεξάρτητη Παραμετροποίηση

Για να δημιουργήσετε ένα λειτουργικό SDS, θα πρέπει να αφιερώσετε πολύ χρόνο για να το σχεδιάσετε. Αυτό σημαίνει πολλά για μια εταιρεία που δεν είναι γεμάτη από ερευνητές που δημιουργούν κάτι παρόμοιο μπορεί να αποδειχθεί μια αρκετά δύσκολη εργασία. Πρέπει λοιπόν να φτιάξουμε κάτι που να μπορούν να χρησιμοποιήσουν όλοι. Σε αυτό το πνεύμα δημιουργούμε συστήματα Independent Domain προκειμένου να δημιουργήσουμε μια σταθερή βάση, ώστε να μπορούν να διευκολύνουν την εκμάθηση μιας πολιτικής με αφηρημένο τρόπο. Οι πολιτικές που μαθαίνονται σε αυτόν τον χώρο βάσης σταθερών διαστάσεων μπορούν να μεταφερθούν σε νέους τομείς. Αυτό γίνεται με τη διερεύνηση της φύσης και της κοινότητας των υποκείμενων εργασιών σε διαφορετικούς τομείς και την παραμετροποίηση των διαφορετικών χρονοθυρίδων ανάλογα με τις σχέσεις και τις πιθανές συνεισφορές τους.

Οι στατιστικές προσεγγίσεις στα Συστήματα Προφορικού Διαλόγου (SDS), ιδιαίτερα, στις Μερικώς Παρατηρήσιμες Διαδικασίες Αποφάσεων Markov (POMDPs), έχουν επιδείξει μεγάλη επιτυχία στη βελτίωση της ευρωστίας των πολιτικών διαλόγου στην επιρρεπή σε σφάλματα Αυτόματη Αναγνώριση Ομιλίας (ASR). Ωστόσο, η δημιουργία στατιστικών SDS (SSDS) για διαφορετικούς τομείς εφαρμογών είναι χρονοβόρα. Παραδοσιακά, κάθε στοιχείο τέτοιων SSDS πρέπει να εκπαιδεύεται με βάση δεδομένα για συγκεκριμένο τομέα, τα οποία δεν είναι πάντα εύκολο να αποκτηθούν. Επιπλέον, σε πολλές περιπτώσεις, θα χρειαστεί να δημιουργηθεί ένα βασικό (π.χ. βάσει κανόνων)

λειτουργικό SDS πριν ξεκινήσει η διαδικασία συλλογής δεδομένων, όπου η ανάπτυξη του αρχικού συστήματος για έναν νέο τομέα απαιτεί σημαντική ποσότητα ανθρώπινης τεχνογνωσίας

Το POMDP είναι ένα ισχυρό εργαλείο για τη μοντελοποίηση διαδοχικών προβλημάτων λήψης αποφάσεων σε συνθήκες αβεβαιότητας, βελτιστοποιώντας την πολιτική για τη μεγιστοποίηση των μακροπρόθεσμων σωρευτικών ανταμοιβών. Συνήθως, σε κάθε στροφή ενός διαλόγου, ένα τυπικό POMDP-SDS αναλύει μια παρατηρούμενη λίστα ASR n-best με βαθμολογίες εμπιστοσύνης σε σημασιολογικές αναπαραστάσεις και εκτιμά μια κατανομή σε στόχους χρήστη, που ονομάζεται κατάσταση πεποίθησης. Μετά από αυτό, η πολιτική διαλόγου επιλέγει μια ενέργεια συστήματος σε σημασιολογικό επίπεδο (ώστε να μας καταλάβει ο οποιοσδήποτε χρήστης), η οποία θα πραγματοποιηθεί από το Natural Language Generation (NLG) πριν από τη σύνθεση της απόκρισης ομιλίας στον χρήστη.

Οι σημασιολογικές αναπαραστάσεις στο SDS αποτελούνται συνήθως από δύο μέρη, μια συνάρτηση επικοινωνίας (π.χ. ενημέρωση, άρνηση, επιβεβαίωση, κ.λπ.) και μια λίστα ζευγών χρονοθυρίδων-τιμών (π.χ. φαγητό=πίτσα, περιοχή=κέντρο, κ.λπ.). Η προηγούμενη γνώση που καθορίζει τις θέσεις-τιμές σε έναν συγκεκριμένο τομέα ονομάζεται οντολογία τομέα. Η βελτιστοποίηση της πολιτικής διαλόγου μπορεί να επιλυθεί μέσω της Ενισχυτικής Μάθησης (RL), όπου ο στόχος είναι να εκτιμηθεί μια ποσότητα Q(b, a), για κάθε b και a, που αντικατοπτρίζει τις αναμενόμενες σωρευτικές ανταμοιβές του συστήματος που εκτελεί την ενέργεια a στην κατάσταση πεποίθησης b, έτσι ώστε η βέλτιστη δράση a να μπορεί να προσδιοριστεί για ένα δεδομένο b σύμφωνα με το a = arg maxa Q(b, a).

Λόγω του εκθετικά μεγάλου χώρου δράσης κατάστασης που μπορεί να προκύψει ένα SDS, είναι απαραίτητη η προσέγγιση της συνάρτησης, όπου υποτίθεται ότι $Q(b, a) \approx f_\theta(\varphi(b, a))$. Εδώ το θ υποδηλώνει την παράμετρο μοντέλου που πρέπει να μαθευτεί και το $\varphi(\cdot)$ είναι μια συνάρτηση χαρακτηριστικών που αντιστοιχίζει το (b, a) σε ένα διάνυσμα χαρακτηριστικών. Για τον υπολογισμό του Q(b, a), μπορεί κανείς να χρησιμοποιήσει είτε μια πεποίθηση περίληψης χαμηλής διάστασης είτε την ίδια την πλήρη πεποίθηση εάν εφαρμοστούν μέθοδοι πυρήνα. Αλλά και στις δύο περιπτώσεις, η ενέργεια «a» θα είναι μια συνοπτική ενέργεια

Οι Zhuoran Wang et al (2013) στην εργασία τους εισήγαγαν μια νέα προσέγγιση για την εξάλειψη της εξάρτησης του τομέα των αναπαραστάσεων κατάστασης και δράσης του διαλόγου, έτσι ώστε οι πολιτικές διαλόγου που εκπαιδεύονται με βάση την προτεινόμενη αναπαράσταση να μπορούν να μεταφερθούν σε διαφορετικούς τομείς. Στα πειράματά

τους, η προτεινόμενη μέθοδος παραμετροποίησης ανεξάρτητου τομέα (DIP) ενσωματώθηκε με έναν γενικό ανιχνευτή κατάστασης διαλόγου για να δώσει έναν συνολικό διαχειριστή διαλόγου ανεξάρτητου τομέα.

Πρώτον, εκπαίδευσαν πολιτικές διαλόγου DIP στον τομέα αναζήτησης εστιατορίων χρησιμοποιώντας GP-SARSA με βάση έναν υπερσύγχρονο προσομοιωτή χρήστη που βασίζεται σε ατζέντα σε σύγκριση με τη διαδικασία εκμάθησης GP-SARSA για το γνωστό σύστημα BUDS. Μετά από αυτό, ανέπτυξαν απευθείας τις πολιτικές DIP που εκπαιδεύτηκαν στον τομέα αναζήτησης εστιατορίων στον τομέα πώλησης φορητών υπολογιστών και συνέκριναν την απόδοσή του με μια πολιτική εντός τομέα που εκπαιδεύτηκε χρησιμοποιώντας τον προσομοιωτή. Τέλος, επιλέγουν τις καλύτερες πολιτικές DIP εντός τομέα και μεταβίβασης και τις ανέπτυξαν σε SDS πώλησης φορητών υπολογιστών από άκρο σε άκρο, για πειράματα με ανθρώπινο θέμα με βάση το MTurk.

Τα πειραματικά αποτελέσματα δείχνουν ότι όταν μεταφέρονται σε έναν νέο τομέα, οι πολιτικές διαλόγου που εκπαιδεύονται με βάση τις αναπαραστάσεις DIP μπορούν να επιτύχουν πολύ κοντινή απόδοση σε αυτές τις πολιτικές που έχουν βελτιστοποιηθεί χρησιμοποιώντας διαλόγους εντός τομέα. Η γεφύρωση του (πολύ μικρού) χάσματος απόδοσης σε αυτήν την περίπτωση θα πρέπει επίσης να είναι απλή, εάν κάποιος λάβει τη μεταφερόμενη πολιτική ως προηγούμενη και πραγματοποιήσει προσαρμογή τομέα.

2.2.4 Δυαδική απεικόνιση χώρου καταστάσεων

Συγκεκριμένα, η προτεινόμενη αναπαράσταση BinLin περιλαμβάνει τα δυαδικά αφηρημένα χαρακτηριστικά για τις υποδοχές τομέα συνδυασμένα γραμμικά. Πιο συγκεκριμένα, οι υποδοχές μετατρέπονται σε ένα γραμμικό δυαδικό διάνυσμα μεγέθους ίσου με τον αριθμό των υποδοχών και σημειώνουμε την παρουσία ή την απουσία τιμής υποδοχής σε κάθε στροφή διαλόγου με μια δυαδική τιμή (0-απών ή 1-παρούσα). Με αυτόν τον τρόπο, το αφηρημένο δυαδικό διάνυσμα χαρακτηριστικών αποτελείται μόνο από n στοιχεία για n υποδοχές. Αυτή η απλή αναπαράσταση μπορεί να επαυξηθεί με ένα βοηθητικό εξάρτημα που περιέχει τις πληροφορίες για το εάν έχει ζητηθεί ή όχι μια υποδοχή από τον χρήστη, πληροφορίες οι οποίες διαφορετικά απορρίπτονται στο BinLin, αν και περιέχονται στην κατάσταση πεποίθησης ρητά. Η προσθήκη του βοηθητικού τμήματος στο BinLin οδηγεί στα χαρακτηριστικά BinAux και προσθέτει στιβαρότητα στην καθαρή δυαδική (BinLin) αναπαράσταση.

Μια οικογένεια νέων αναπαραστάσεων καταστάσεων που βασίζονται σε δυαδικά χαρακτηριστικά παρουσιάζεται αρχικά σε αυτό το άρθρο. Αυτή η αναπαράσταση είναι

μεταβιβάσιμη στον τομέα και συμπαγής, αλλά στιβαρή. Η βασική ιδέα είναι ότι, σε πολλούς τομείς, η γνώση των ακριβών τιμών υποδοχής μπορεί να μην είναι τόσο σημαντική για την επιλογή ενεργειών διαλόγου, όσο η γνώση του εάν μια τιμή υποδοχής είναι γνωστή ή όχι. Επομένως, τα προτεινόμενα χαρακτηριστικά σηματοδοτούν την παρουσία ή την απουσία μιας τιμής υποδοχής σε κάθε στροφή διαλόγου με μια δυαδική τιμή (0-απών ή 1-παρούσα). Αυτά τα αφηρημένα χαρακτηριστικά συνδυάζονται γραμμικά για όλες τις υποδοχές τομέα (BinLin) και μπορούν να επαυξηθούν με βοηθητικές δυαδικές πληροφορίες σχετικά με αιτήματα υποδοχής από τον χρήστη (BinAux). Η πιο κοινή προσέγγιση για την επίλυση του προβλήματος βελτιστοποίησης της διαδοχικής λήψης αποφάσεων είναι η χρήση της ενισχυτικής μάθησης (RL).

Πιο πρόσφατα, αξιοποιείται η χρήση των βαθιών νευρωνικών δικτύων (NN) για την επίλυση του προβλήματος βελτιστοποίησης. Αυτό μπορεί να αποδοθεί εν μέρει στο γεγονός ότι οι αρχιτεκτονικές σε βάθος με πολλά κρυφά επίπεδα μπορούν να χρησιμοποιηθούν αποτελεσματικά για πολύπλοκες εργασίες και περιβάλλοντα. Πολλά υποσχόμενα αποτελέσματα έχουν επιτευχθεί για παράδειγμα με τα συστήματα Deep Q-Network (DQN).

Συμφωνώντας με τους Margarita Kotti et al (2018: 2) που συνέκρινε την κύρια χρήση του υπολογισμού A2C με συμπαγείς αναπαραστάσεις (Plunge, BinLin, BinAux, η μείωση των μετρήσεων που ελήφθησαν από τη μετατροπή του BS σε Plunge ή Bin{Lin,Aux} έχει ουσιαστικά σημεία ενδιαφέροντος, τόσο όταν η προετοιμασία και η δοκιμή του πλαισίου γίνονται κάτω από τις ίδιες συνθήκες σημασιολογικού ποσοστού σφαλμάτων, όσο και όταν το ποσοστό σημασιολογικού σφάλματος δοκιμής είναι σημαντικά υψηλότερο από το προπαρασκευαστικό. Αυτό μπορεί να πιστωθεί κάπως στην επανάληψη και πενιχρή κατάσταση της τυπικής αναπαράστασης BS Τα πρόσφατα προτεινόμενα σημεία του BinAux είναι i) μεταβιβάσιμος τομέας, ii) χαμηλής διάστασης, που διαφημίζει μια διάταξη μείωσης μεγέθους σε σύγκριση με το συνολικό BS, iii) προσαρμόσιμη σε χώρους με πολλούς χώρους και πολλές τιμές ανοίγματος , και iv) υπολογιστικά αποδοτικό.

2.2.5 Κωδικοποίηση χώρου από αυτόματους-κωδικοποιητές

Τα AE είναι μια οικογένεια τοπολογιών NN που χρησιμοποιούνται για μάθηση χωρίς επίβλεψη και έχουν χρησιμοποιηθεί επιτυχώς για μη γραμμική εξαγωγή χαρακτηριστικών σε διάφορους τομείς εφαρμογών. Λόγω της αρχιτεκτονικής τους, οι AE

αναγκάζονται να μάθουν τις χαμηλότερες διαστάσεις και τις πιο ισχυρές αναπαραστάσεις του διανύσματος εισόδου x. Τυπικά, τα AEs μπορούν να θεωρηθούν ως ο συνδυασμός δύο δικτύων. Ο πρώτος είναι ο κωδικοποιητής, ο οποίος παίρνει το διάνυσμα εισόδου σε έναν λανθάνοντα χώρο χαμηλότερης διάστασης. Το δεύτερο είναι ο αποκωδικοποιητής, ο οποίος παίρνει την κωδικοποιημένη αναπαράσταση και την μεταφέρει πίσω στην αρχική είσοδο. Η αρχιτεκτονική τους παρουσιάζει τέλεια συμμετρία ως προς το κεντρικό κρυφό στρώμα που είναι το χαμηλότερου διαστάσεων και χρησιμεύει ως στρώμα κωδικοποίησης. (Lygerakis, June 2019)

Ο Fotios Lygerakis et al (2019) εισήγαγε μια νέα χρήση των Autoencoders (AEs). Ο στόχος εδώ ήταν να αποκτήσουμε μια χαμηλών διαστάσεων, σταθερού μήκους και συμπαγή, αλλά στιβαρή αναπαράσταση του χώρου BS. Έχει διερευνηθεί η χρήση πυκνών AE, Denoising AE (DAE) και Variational Denoising AE (VDAE), τα οποία συνδύασαν με το GP-SARSA για την εκμάθηση πολιτικών διαλόγου στην εργαλειοθήκη PyDial. Σε αυτό το πλαίσιο, το BS αναπαρίσταται κανονικά σε έναν σχετικά συμπαγή, αλλά ακόμα περιττό χώρο περίληψης, ο οποίος λαμβάνεται μέσω μιας ευρετικής χαρτογράφησης του αρχικού κύριου χώρου. (Lygerakis, June 2019)

Οι κωδικοποιητές είναι μια οικογένεια τοπολογιών NN που χρησιμοποιούνται για μάθηση χωρίς επίβλεψη και έχουν χρησιμοποιηθεί αποτελεσματικά για μη γραμμική εξαγωγή επισήμανσης σε λίγους χώρους εφαρμογής. Λόγω της μηχανικής τους, τα AE περιορίζονται να απομνημονεύουν χαμηλότερες διαστάσεις και πιο ισχυρές αναπαραστάσεις του εισερχόμενου διανύσματος x. Τακτικά, τα AEs μπορούν να θεωρηθούν ως η συνένωση δύο συστημάτων. Ο κύριος είναι ο κωδικοποιητής, ο οποίος μεταφέρει το στρώμα εισόδου σε έναν αδρανή χώρο χαμηλότερης διάστασης. Η στιγμή που κάποιος είναι ο αποκωδικοποιητής, ο οποίος παίρνει την κωδικοποιημένη συμπαγή αναπαράσταση και την επαναφέρει στην αρχική είσοδο. Η μηχανική τους δείχνει μια εξιδανικευμένη συμμετρία ως προς το κεντρικό καλυμμένο στρώμα που είναι το χαμηλότερου διαστάσεων και χρησιμεύει ως το στρώμα κωδικοποίησης. (Lygerakis, June 2019)

2.2.6 Φεουδαρχικά Χαρακτηριστικά

Το βασικό χαρακτηριστικό της φεουδαρχικής αρχιτεκτονικής είναι εν μέρει μια από τα κύρια προβλήματα στην ενίσχυση της εκμάθησης σχετικά με τον τρόπο διαίρεσης μιας μεμονωμένης εργασίας σε επιμέρους εργασίες σε πολλαπλά επίπεδα. Καταλαβαίνουμε

απο μια επίδειξη του πώς αυτό μπορεί να γίνει χωριστά από την επιλογή μεταξύ διαφορετικών πιθανών υπο-διαχειριστών σε ένα δεδομένο επίπεδο. Εξαρτάται από το αν υπάρχει ένα λογικό, διαχειριστικό και χρήσιμο σύστημα, κατά προτίμηση βασισμένο σε μια φυσική ιεραρχική διαίρεση του διαθέσιμου χώρου κατάστασης. Για ορισμένες εργασίες μπορεί να είναι πολύ αναποτελεσματικό, καθώς αναγκάζει κάθε υποδιευθυντή να μάθει πώς να ικανοποιεί όλες τις δευτερεύουσες εργασίες που έχει ορίσει ο διευθυντής του, ανεξάρτητα από το αν αυτές οι δευτερεύουσες εργασίες είναι κατάλληλες. Επομένως, είναι πιο πιθανό να είναι χρήσιμο σε περιβάλλοντα στα οποία μπορούν να αλλάξουν οι καθορισμένες εργασίες. Οι διευθυντές δεν χρειάζεται απαραίτητα να γνωρίζουν εκ των προτέρων τις συνέπειες των πράξεών τους. Θα μπορούσαν να μάθουν, με αυτό-εποπτευόμενο τρόπο, πληροφορίες σχετικά με τις πολιτειακές μεταβάσεις που έχουν βιώσει. Αυτές οι παρατηρούμενες επόμενες καταστάσεις μπορούν να χρησιμοποιηθούν ως στόχοι για τους υπό-διαχειριστές τους - η συνέπεια στην παροχή ανταμοιβών για τις κατάλληλες μεταβάσεις είναι η μόνη απαίτηση.

Η ενισχυτική μάθηση (RL) είναι μια πολλά υποσχόμενη προσέγγιση για την επίλυση της βελτιστοποίησης της πολιτικής διαλόγου. Οι παραδοσιακοί αλγόριθμοι RL, ωστόσο, αποτυγχάνουν να κλιμακωθούν σε μεγάλους τομείς λόγω της κατάρας της διάστασης.

Οι Inigo˜ Casanueva et al (2015) προτείνουν μια νέα αρχιτεκτονική διαχείρισης διαλόγου, βασισμένη στο Feudal RL, η οποία αποσυνθέτει την απόφαση σε δύο βήματα. ένα πρώτο βήμα όπου μια κύρια πολιτική επιλέγει ένα υποσύνολο πρωτόγονων ενεργειών και ένα δεύτερο βήμα όπου μια πρωταρχική ενέργεια επιλέγεται από το επιλεγμένο υποσύνολο. Έδειξαν ότι μια εφαρμογή αυτής της προσέγγισης, βασισμένης στα δίκτυα Deep-Q, ξεπερνά σημαντικά την προηγούμενη στάθμη της τέχνης σε πολλούς τομείς διαλόγου και περιβάλλοντα, χωρίς την ανάγκη πρόσθετου σήματος ανταμοιβής.

Σε διαλόγους πλήρωσης θυρίδων, μια μέθοδος HRL που βασίζεται εκ νέου στην αφαίρεση χώρου, όπως το Feudal RL (FRL), θα πρέπει να επιτρέπει την κλίμακα RL σε τομείς με μεγάλο αριθμό θυρίδων. Το FRL διαιρεί μια εργασία χωρικά και όχι χρονικά, αποσυνθέτοντας τις αποφάσεις σε πολλά βήματα και χρησιμοποιώντας διαφορετικά επίπεδα αφαίρεσης σε κάθε υπό-απόφαση. Αυτό το πλαίσιο είναι ιδιαίτερα χρήσιμο σε εργασίες RL με μεγάλους διακριτούς χώρους δράσης, καθιστώντας το πολύ ελκυστικό για τη διαχείριση διαλόγου μεγάλων τομέων

Οι Inigo˜ Casanueva et al (2015) παρουσίασαν μια Πρωτόγονη Προσέγγιση Λόγου που αναλύει την επιλογή σε κάθε στροφή σε δύο βήματα. Σε ένα πρώτο βήμα, η προσέγγιση

επιλέγει τυχαία ότι χρειάζεται μια εναρκτήρια αυτόνομη ή διαστημική δευτερεύουσα δραστηριότητα. Σε εκείνο το σημείο, η κατάσταση κάθε υπο-πολιτικής χώρου είναι ονειρική για να ληφθούν υπόψη τα κυριότερα σημεία που σχετίζονται με αυτόν τον χώρο και επιλέγεται μια πρωτόγονη δραστηριότητα από το ήδη επιλεγμένο υποσύνολο. Το FRL αναλύει την επιλογή διάταξης (b) = a σε κάθε στροφή σε μερικές επιμέρους αποφάσεις, χρησιμοποιώντας διαφορετικά ονειρικά μέρη της κατάστασης πεποίθησης σε κάθε υπό-απόφαση. Ο στόχος ενός προγραμματισμένου SDS ανάθεσης είναι η εκπλήρωση του στόχου των πελατών, αλλά καθώς ο στόχος δεν είναι γνωστός για το SDS, το SDS θα πρέπει να συγκεντρώνει αρκετά δεδομένα για την ακριβή εκπλήρωσή του.

Επομένως, σε κάθε στροφή, το DM μπορεί να χωρίσει την απόφασή του σε δύο βήματα: πρώτον, να αποφασίσει μεταξύ της ανάληψης μιας ενέργειας για τη συλλογή πληροφοριών σχετικά με τον στόχο του χρήστη (ενέργειες συλλογής πληροφοριών) ή της λήψης μιας ενέργειας για την εκπλήρωση του στόχου χρήστη ή ενός μέρους του (ενέργειες που παρέχουν πληροφορίες) και δεύτερον, επιλέξτε μια ενέργεια για εκτέλεση από το προηγουμένως επιλεγμένο υποσύνολο. Σε έναν διάλογο συμπλήρωσης θυρίδων, το σύνολο των ενεργειών συλλογής πληροφοριών μπορεί να οριστεί ως το σύνολο των ενεργειών που εξαρτώνται από τη θέση υποδοχής, ενώ το σύνολο των ενεργειών παροχής σε σχηματισμό μπορεί να οριστεί ως οι υπόλοιπες ενέργειες.

Η διαχείριση του φεουδαρχικού διαλόγου αποσυνθέτει την πολιτική απόφαση π(b) =a σε κάθε στροφή σε πολλές υπό-αποφάσεις, χρησιμοποιώντας διαφορετικά αφηρημένα μέρη της κατάστασης πεποίθησης σε κάθε υπό-απόφαση. Ο στόχος ενός task-oriented SDS είναι να εκπληρώσει τον στόχο των χρηστών, αλλά καθώς ο στόχος δεν είναι παρατηρήσιμος για το SDS, το SDS πρέπει να συγκεντρώσει αρκετές πληροφορίες για να τον εκπληρώσει σωστά. Επομένως, σε κάθε στροφή, το DM μπορεί να πάρει 2 διαφορετικές κατευθύνσεις: πρώτον, να αποφασίσει μεταξύ της ανάληψης μιας ενέργειας για τη συλλογή πληροφοριών σχετικά με τον στόχο του χρήστη (ενέργειες συλλογής πληροφοριών) ή της λήψης μιας ενέργειας για την εκπλήρωση του στόχου χρήστη ή ενός μέρους του (ενέργειες που παρέχουν πληροφορίες) και δεύτερον, επιλέγει μια (πρωτόγονη) ενέργεια για εκτέλεση από το προηγουμένως επιλεγμένο υποσύνολο. Σε έναν διάλογο συμπλήρωσης χρονοθυρίδων, το σύνολο των ενεργειών συλλογής πληροφοριών μπορεί να οριστεί ως το σύνολο των ενεργειών που εξαρτώνται από την υποδοχή, ενώ το σύνολο των ενεργειών που παρέχουν πληροφορίες μπορεί να οριστεί ως οι υπόλοιπες ενέργειες.

Η ενισχυτική μάθηση (RL) είναι μια πολλά υποσχόμενη προσέγγιση για την επίλυση της βελτιστοποίησης της πολιτικής του διαλόγου. Οι παραδοσιακοί αλγόριθμοι RL, ωστόσο, αποτυγχάνουν να κλιμακωθούν σε μεγάλους τομείς λόγω της διάστασής τους. Προτείνουμε μια νέα αρχιτεκτονική διαχείρισης διαλόγου, βασισμένη στο Feudal RL, η οποία αποσυνθέτει την απόφαση σε δύο βήματα: ένα πρώτο βήμα όπου μια κύρια πολιτική επιλέγει ένα υποσύνολο πρωτόγονων ενεργειών και ένα δεύτερο βήμα όπου μια πρωταρχική ενέργεια επιλέγεται από το επιλεγμένο υποσύνολο. Οι δομικές πληροφορίες που περιλαμβάνονται στην οντολογία τομέα χρησιμοποιούνται για την αφαίρεση του χώρου καταστάσεων διαλόγου, λαμβάνοντας τις αποφάσεις σε κάθε βήμα χρησιμοποιώντας διαφορετικά μέρη της αφηρημένης κατάστασης. Αυτό, σε συνδυασμό με έναν μηχανισμό ανταλλαγής πληροφοριών μεταξύ των υποδοχών, αυξάνει την επεκτασιμότητα σε μεγάλους τομείς.

2.3 Αυτό-ενισχυτική Μάθηση (Reinforcement Learning)

Οι τεχνικές Ενισχυτικής Μάθησης χρησιμοποιούνται για τον καθορισμό μιας πολιτικής βάσει της οποίας ο διάλογος λαμβάνει μια απόφαση με τη μορφή δράσης. Σε αυτό το πλαίσιο, το σύστημα μαθαίνει μέσω μιας διαδικασίας δοκιμής και σφάλματος που διέπεται από ένα δυνητικά καθυστερημένο σήμα ανταμοιβής. Επομένως, η ενότητα DM μαθαίνει να σχεδιάζει ενέργειες προκειμένου να μεγιστοποιήσει το τελικό αποτέλεσμα. Η εκμάθηση ενός πράκτορα RL περιλαμβάνει διαδοχική παρατήρηση και ανατροφοδότηση που δημιουργείται από τον λειτουργικό του κόσμο. Αυτή η ανατροφοδότηση εξαρτάται από τη δράση του ίδιου του πράκτορα. Επομένως, δύο διαφορετικές πολιτικές θα δημιουργήσουν δύο διαφορετικές ακολουθίες παρατηρήσεων. Έχουν προταθεί πολιτικές εκπαίδευσης και δοκιμών που αλληλεπιδρούν άμεσα με πραγματικούς χρήστες. Ωστόσο, η πολυπλοκότητα του συστήματος, ο χρόνος και το υψηλό κόστος καθιστούν αυτή την προσέγγιση ανέφικτη για ένα μεγάλο μέρος της ερευνητικής κοινότητας. Επιπλέον, μπορεί να είναι πολύ δύσκολο να ελεγχθεί για εξωγενείς παράγοντες που μπορούν να τροποποιήσουν τη συμπεριφορά των χρηστών, όπως η διάθεση ή η κούραση, καθιστώντας μια δίκαιη αξιολόγηση πολύ δύσκολη.

Οι Iñigo Casanueva et al (2017) απέδειξαν ότι ένας αριθμός παραμετρικών υπολογισμών πράκτορα, ειδικά οι υπολογισμοί μάθησης βαθιάς υποστήριξης - DQN, A2C και Common Actor-Critic σε σύγκριση με μια μη παραμετρική επίδειξη, το GP-SARSA μπορεί να δημιουργήσει ένα σύστημα δοκιμαστικής βάσης για εκ των προτέρων δοκιμών και να ενθαρρύνει τη διερευνητική δημιουργικότητα. Οι συνεργάτες του Exchange

γίνονται γρήγορα μια κρίσιμη βοήθεια μέρα με τη μέρα. Για να αποφύγετε την αξιοσημείωτη προσπάθεια που απαιτείται για τη δημιουργία της επιθυμητής ροής λόγου, η ενότητα Διαχείρισης Λόγου (DM) μπορεί να μετατραπεί ως αδιάκοπη προετοιμασία Markov (MDP) και να προετοιμαστεί μέσω του Fortification Learning (RL).

Πολλά μοντέλα RL έχουν διερευνηθεί τα τελευταία χρόνια. Ωστόσο, η έλλειψη κοινού πλαισίου συγκριτικής αξιολόγησης καθιστά δύσκολη τη δίκαιη σύγκριση μεταξύ διαφορετικών μοντέλων και την ικανότητά τους να γενικεύονται σε διαφορετικά περιβάλλοντα.

2.3.1 Gaussian Process (GP-SARSA)

Το SARSA είναι μια αρκετά απλή επέκταση του αλγορίθμου TD, στον οποίο υπολογίζονται οι τιμές των ενεργειών κατάστασης, επιτρέποντας έτσι την εκτέλεση βημάτων βελτίωσης πολιτικής χωρίς να απαιτείται πρόσθετη γνώση για το μοντέλο MDP. Η ιδέα είναι να χρησιμοποιηθεί η σταθερή πολιτική μ που ακολουθείται προκειμένου να οριστεί μια νέα, επαυξημένη διαδικασία, διατηρώντας το ίδιο μοντέλο ανταμοιβής. Ο ίδιος συλλογισμός μπορεί να εφαρμοστεί για να εξαχθεί ένας αλγόριθμος GP SARSA από τον αλγόριθμο GPTD. Το μόνο που χρειαζόμαστε είναι να ορίσουμε μια συνάρτηση πυρήνα συνδιακύμανσης σε ζεύγη κατάστασης-δράσης, δηλ. k: (X ×U)×(X ×U)→R. Επειδή οι καταστάσεις και οι ενέργειες είναι διαφορετικές οντότητες, είναι λογικό να αποσυντεθεί το k σε έναν πυρήνα κατάστασης kx και έναν πυρήνα δράσης ku:k(x,u,x′,u′) =kx(x,x′)ku(u, u').

Το μόνο που μένει τώρα είναι να εκτελέσετε το GPTD στην ακολουθία επαυξημένης ανταμοιβής κατάστασης, χρησιμοποιώντας τη νέα συνάρτηση πυρήνα κατάστασης δράσης.

2.3.2 Least-Squares Policy Iteration

Το LSPI υποθέτει ότι η συνάρτηση Q μπορεί να αναπαρασταθεί ως ένα σταθμισμένο άθροισμα των χαρακτηριστικών φi : $Q(b,\hat{a}) = \sum_{i=1}^{k} \varphi_i(b,\hat{a})\theta_i = \varphi(b,\hat{a})^T\theta$ όπου θ είναι ένα διάνυσμα βάρους.Τα χαρακτηριστικά εξόδου του b που είναι σημαντικά για τη λήψη αποφάσεων που σχετίζονται με τη συνοπτική ενέργεια â και τα θ και φ(b,â) είναι και τα δύο μεγέθους κ.

2.3.3 Deep Q-Network DQN

Όσον αφορά το τμήμα του διευθυντή διακανονισμού, οι συμβατικές προσεγγίσεις χρησιμοποιούν είτε τα MDPs είτε το εν μέρει POMPDS για να κατανοήσουν το ζήτημα της διαδοχικής επιλογής. Η πιο κοινή προσέγγιση για την κατανόηση του ζητήματος βελτιστοποίησης της λήψης διαδοχικών επιλογών είναι η χρήση της RL. Τα περισσότερα πρόσφατα, η χρήση των βαθιών νευρικών συστημάτων (NNs) για την κατανόηση του ζητήματος της βελτιστοποίησης γίνεται κατάχρηση. Αυτό μπορεί να πιστωθεί ως επί το πλείστον στην αλήθεια ότι τα βαθιά μοντέλα με μερικά κρυφά επίπεδα μπορούν να χρησιμοποιηθούν επάρκεια για περίπλοκες εργασίες και καταστάσεις.

Πολλά υποσχόμενα αποτελέσματα έχουν επιτευχθεί, για παράδειγμα, με τα συστήματα Deep Q-Network (DQN), τα οποία οδηγούν σε πολλές παραλλαγές, όπως το NDQN. Πρόσφατα, εμφανίστηκε μια τάση προς μεθόδους διαβάθμισης πολιτικής, όπως το Advantage Actor-Critic (A2C), οι οποίες έχουν αποδειχθεί αποτελεσματικές σε παιχνίδια Atari, προσομοιωτές αυτοκινήτου και προσομοιωτές φυσικής. Εστιάζοντας στα συστήματα διαλόγου, οι συγγραφείς δοκιμάζουν έναν βαθύ αλγόριθμο A2C, είτε αρχικοποιημένο με εποπτευόμενη μάθηση είτε όχι, στον τομέα των εστιατορίων, όπως σε αυτή την εργασία. Το βασικό τους εύρημα είναι ότι ο αλγόριθμος A2C συνέκλινε γρηγορότερα από τους αλγόριθμους DQN και GP-SARSA

.

2.3.4 ENAC (Episodic Natural Actor Critics)

Οι στρατηγικές RL με κριτές είναι διαδικτυακές προσεγγίσεις στον κύκλο διευθέτησης στον οποίο οι παράμετροι της εργασίας εκτίμησης αξιολογούνται χρησιμοποιώντας την κοσμική μάθηση διάκρισης και οι παράμετροι προσέγγισης αναθεωρούνται με στοχαστική πτώση κλίσης. Οι στρατηγικές που βασίζονται σε γωνίες προσέγγισης με αυτόν τον τρόπο είναι εξαιρετικά ενδιαφέρουσες λόγω της συμβατότητάς τους με τις στρατηγικές εκτίμησης εργασίας, οι οποίες απαιτούνται για τον χειρισμό τεράστιων ή ατέρμονων χώρων καταστάσεων. Η χρήση της κοσμικής μάθησης διάκρισης με αυτόν τον τρόπο είναι ασυνήθιστη ενδιαφέρουσα, καθώς σε πολλές εφαρμογές μειώνει σημαντικά τις διακυμάνσεις των μετρητών κλίσης. Η χρήση της κοινής γωνίας είναι περίεργη καθώς μπορεί να δημιουργήσει ανώτερες εξαρτημένες παραμετροποιήσεις και έχει φανεί ότι μειώνει τις διακυμάνσεις σε λίγες περιπτώσεις.

2.3.5 AAC: Advanced Actor Critics

Τα τελευταία χρόνια, μια σειρά από ερευνητικές ομάδες έχουν διερευνήσει τη χρήση μιας εγκατάστασης που βασίζεται σε προσομοίωση δύο σταδίων. Ένα στατιστικό μοντέλο χρήστη εκπαιδεύεται πρώτα σε περιορισμένο αριθμό δεδομένων διαλόγου και το μοντέλο χρησιμοποιείται στη συνέχεια για την προσομοίωση διαλόγων με τον διαδραστικά μαθησιακό DM (βλ. Schatzmann et al. (2006) για βιβλιογραφική ανασκόπηση).

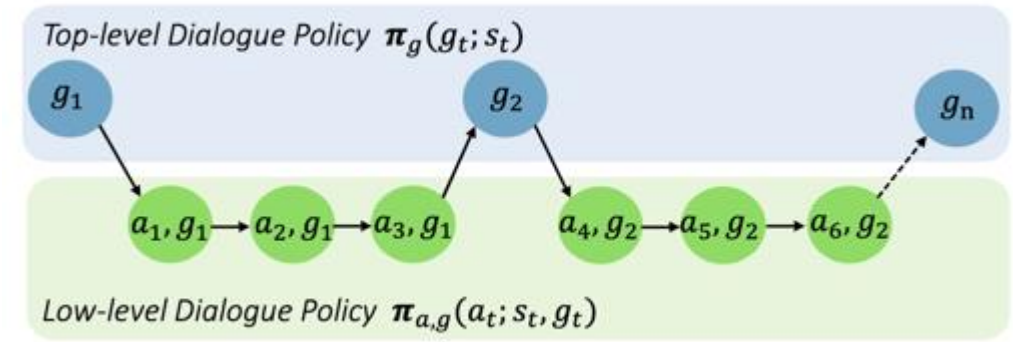
Η προσέγγιση που βασίζεται στην προσομοίωση προϋποθέτει την παρουσία ενός μικρού σώματος από κατάλληλα σχολιασμένους διαλόγους εντός τομέα ή διαλόγους εκτός τομέα με αντίστοιχη μορφή διαλόγου (Lemon et al., 2006). Σε περιπτώσεις που δεν υπάρχουν διαθέσιμα τέτοια δεδομένα, οι χειροποίητες τιμές μπορούν να αντιστοιχιστούν στις παραμέτρους του μοντέλου δεδομένου ότι το μοντέλο είναι αρκετά απλό (Levin et al., 2000; Pietquin and Dutoit, 2005) αλλά η απόδοση των πολιτικών διαλόγου που έχουν μάθει με αυτόν τον τρόπο δεν έχει αξιολογήθηκε χρησιμοποιώντας πραγματικούς χρήστες

.

2.4 Χώρος Δραστηριοτήτων (Action Space)

Για πολύπλοκους διαλόγους, επιλέγεται μια βασική ενέργεια συστήματος σε κάθε βήμα των παραδοσιακών μεθόδων RL, όπως η αναζήτηση της τιμής της υποδοχής ή η επιβεβαίωση περιορισμών. Στη λειτουργία HRL, επιλέγεται ένα σύνολο βασικών ενεργειών με βάση την πολιτική ανώτατου επιπέδου και, στη συνέχεια, επιλέγεται μια βασική ενέργεια από το τρέχον σύνολο με βάση την πολιτική χαμηλού επιπέδου, όπως φαίνεται στην Εικόνα 2.4.

Αυτή η προοδευτική διαίρεση των χώρων δραστηριότητας καλύπτει τους περιορισμούς της χρονικής διευθέτησης μεταξύ διαφορετικών επιμέρους εργασιών, οι οποίοι ενθαρρύνουν την ολοκλήρωση σύνθετων εργασιών. Κατά την επέκταση, η εγγενής αντιστάθμιση ηρεμεί βιώσιμα το ζήτημα των πενιχρών ανταμοιβών, επιταχύνοντας την προετοιμασία του RL, την πρόβλεψη ανταλλαγής λόγου μεταξύ επιμέρους καθηκόντων και την προώθηση της ακρίβειας της πρόβλεψης δραστηριότητας. Φυσικά, το προοδευτικό σχέδιο δραστηριοτήτων απαιτεί βασικές πληροφορίες και τα είδη των δευτερευουσών εργασιών πρέπει να αποφασίζονται από ειδικούς.



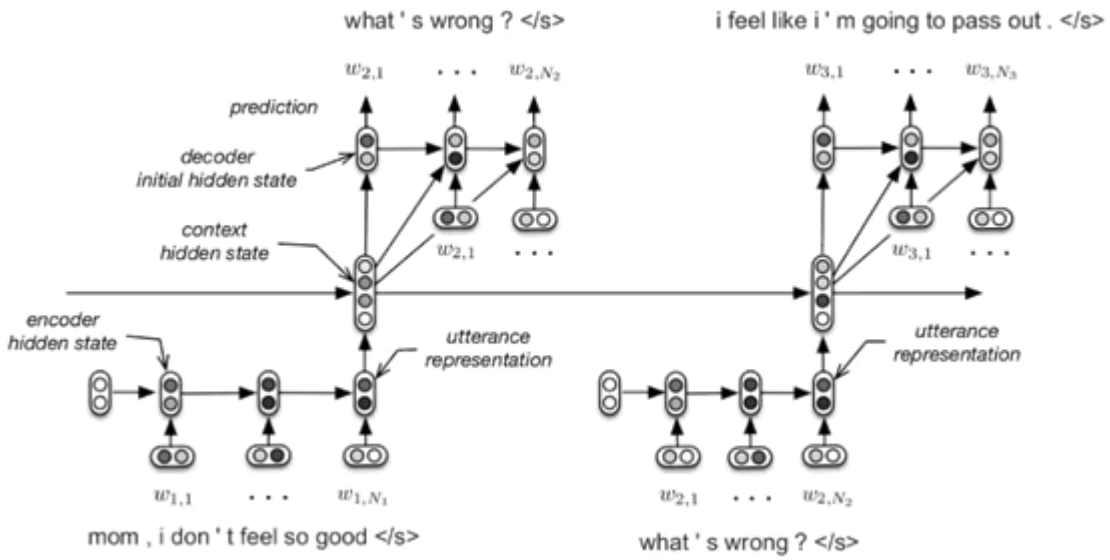
Εικόνα 4:Policy selection process of HRL

2.5 Διαφορικές προσεγγίσεις

Ο Iulian V. Serban et all (2016) εξέτασε τη δουλειά της δημιουργίας ανοιχτού χώρου, πλαισίων λόγου συνομιλίας που βασίζονται σε επεκτατικά σώματα ανταλλαγής χρησιμοποιώντας παραγωγικά μοντέλα. Τα παραγωγικά μοντέλα δημιουργούν αντιδράσεις πλαισίου που δημιουργούνται ανεξάρτητα λέξη προς λέξη, ανοίγοντας την αληθοφάνεια για πρακτικό, προσαρμόσιμο διαισθητικό. Ενίσχυσαν την τελευταία προτεινόμενη διάφορη επαναλαμβανόμενη νευρωνική οργάνωση κωδικοποιητή-αποκωδικοποιητή στο χώρο του λόγου, και δείχνουν ότι αυτή η εκπομπή είναι ανταγωνιστική με μοντέλα νευρωνικής διαλέκτου τελευταίας τεχνολογίας και μοντέλα n-gram back-off.

Η δέσμευσή τους είναι εντός της πορείας πλαισίων που μπορούν να εκπαιδεύσουν από άκρο σε άκρο, που δεν βασίζονται σε στόχους και βασίζονται σε παραγωγικά πιθανοτικά μοντέλα. Χαρακτήρισαν το ζήτημα του γενεσιουργού λόγου ως μοντελοποίηση των αρθρώσεων και της διαισθητικής δομής του λόγου. Ως εκ τούτου, βλέπουν την παράστασή τους ως ένα γνωστικό πλαίσιο, το οποίο χρειάζεται να πραγματοποιήσει χαρακτηριστική διαλεκτική κατανόηση, σκέψη, επιλογή και χαρακτηριστική εποχή διαλέκτου προκειμένου να αναπαράγει ή να μιμηθεί τη συμπεριφορά των χειριστών μέσα στο σώμα που ετοιμάζει. Η προσέγγισή τους διαφέρει από προηγούμενες εργασίες για την εκμάθηση συστημάτων διαλόγου μέσω της αλληλεπίδρασης με ανθρώπους, επειδή μαθαίνει εκτός σύνδεσης μέσω παραδειγμάτων διαλόγων ανθρώπου-ανθρώπου και στοχεύει να μιμηθεί τους διαλόγους στο εκπαιδευτικό σώμα αντί να μεγιστοποιήσει έναν συγκεκριμένο στόχο. λειτουργία. Σε αντίθεση με τη μάθηση που βασίζεται σε εξηγήσεις και τα συστήματα συμπερασμάτων που βασίζονται σε κανόνες (Langley et al. 2014), το

μοντέλο μας δεν απαιτεί μια προκαθορισμένη αναπαράσταση χώρου κατάστασης ή δράσης. Πειραματίστηκαν με τα καθιερωμένα επαναλαμβανόμενα νευρωνικά δίκτυα (RNN) και τα μοντέλα n-gram. Συγκεκριμένα, έχει υιοθετηθεί ο ιεραρχικός επαναλαμβανόμενος κωδικοποιητής-αποκωδικοποιητής (HRED) και έδειξε ότι είναι ανταγωνιστικός με άλλα μοντέλα της βιβλιογραφίας. Αυτοί οι ερευνητές επέκτειναν την αρχιτεκτονική του μοντέλου για να ταιριάζει καλύτερα στην εργασία του διαλόγου.



Εικόνα 5:The computational graph of the HRED architecture for a dialogue composed of three turns

Είχε αποδειχθεί ότι μια διαφορετική επίδειξη δημιουργίας επαναλαμβανόμενης νευρικής οργάνωσης σε επίπεδο επίπεδο μπορεί να υπερτερεί τόσο των μοντέλων που βασίζονται σε n-gram όσο και των μοντέλων νευρικής διάταξης μοτίβων κατά τη μοντελοποίηση αρθρώσεων και πράξεων λόγου. Για να υποστηρίξει αυτή την εξέταση, είχε παρουσιαστεί ένα νέο σύνολο δεδομένων που ονομάζεται MovieTriples βασισμένο σε σενάρια κινηματογραφικών ταινιών, τα οποία είναι λογικά για τη μοντελοποίηση μεγάλου, ανοιχτού χώρου που μιλάει κοντά στην ανθρώπινη ομιλούμενη διάλεκτο. Επεκτείνοντας τον επαναλαμβανόμενο προοδευτικό σχεδιασμό, καθιέρωσαν δύο ζωτικής σημασίας διορθώσεις για την εκτέλεση βημάτων: τη χρήση ενός εκτεταμένου εξωτερικού μονολογικού σώματος για να αρχικοποιήσουν τις ενσωματώσεις λέξεων και τη χρήση ενός τεράστιου σχετικού, αλλά μη διαλόγου, σώματος για τη διευθέτηση της προεκπαίδευσης το επαναλαμβανόμενο δίχτυ. Αυτό εστιάζει στην απαίτηση για μεγαλύτερα σύνολα δεδομένων λόγου.

3 ΠΕΙΡΑΜΑΤΑ

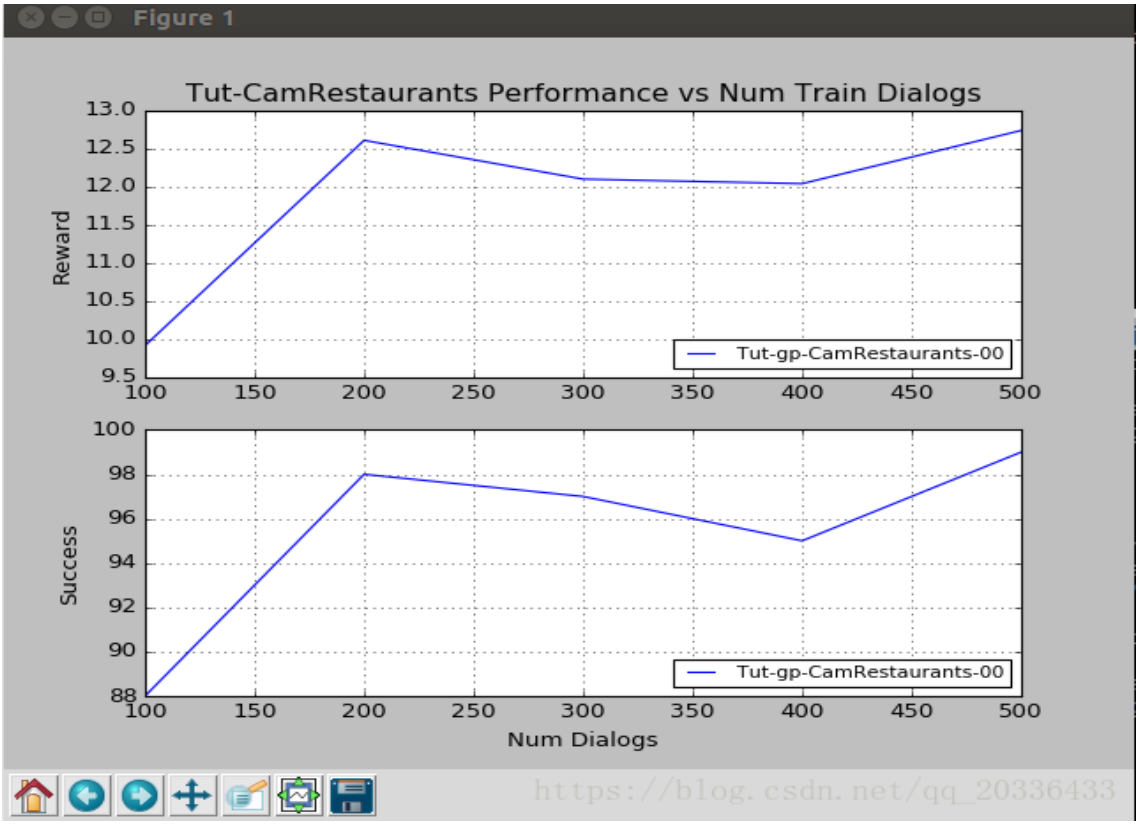
Αυτή η ενότητα περιλαμβάνει πειράματα που κάναμε για να συγκρίνουμε μερικές από τις μεθόδους που περιγράφονται παραπάνω.

3.1 The PyDial toolkit

Η πιο βασική επίδειξη συστήματος στην εργαλειοθήκη χρησιμοποιεί μια πολύ απλή χειρόγραφη και βασισμένη σε κανόνες στρατηγική συνομιλίας. Το Pydial.py παρέχει την πιο βασική λειτουργικότητα για στρατηγικές εκπαίδευσης και δοκιμών.
Κατά την ανάθεση πιθανών καταστάσεων πεποίθησης, η διαχείριση συνομιλιών μπορεί να θεωρηθεί ως πρόβλημα ελέγχου. Μπορούμε να χρησιμοποιήσουμε την ενισχυτική μάθηση για να εκπαιδεύσουμε πολιτικές. Μπορούμε να δημιουργήσουμε μια συνάρτηση ανατροφοδότησης που την ανταμείβει όταν υπάρχει σωστή πολιτική, διαφορετικά την τιμωρεί.
οι απαιτούμενες παράμετροι ορίζονται στο αρχείο διαμόρφωσης Tut-gp-CamRestaurants .cfg, ο συγκεκριμένος μηχανισμός λειτουργίας μπορεί να αναφέρεται στον κώδικα. Η προεπιλεγμένη ρύθμιση είναι να εκπαιδεύεστε σε πέντε παρτίδες, κάθε παρτίδα εκατό ομάδων συνομιλιών με ποσοστό σφάλματος 0. Η ταχύτητα εκπαίδευσης εξαρτάται από την ταχύτητα του υπολογιστή σας και έκανα εκπαίδευση για περίπου οκτώ λεπτά περίπου. Όταν ολοκληρωθεί, θα δημιουργηθεί ένας επιπλέον φάκελος _tutorialpolicies στον κατάλογο, με συνολικά πέντε έως δέκα αρχεία, το κάθε ζεύγος περιέχει το αρχείο

λεξικού σημείων εκπαίδευσης (.dct) και ένα σύνολο ονομάτων αρχείων παραμέτρων (.pcr) που περιέχουν το ποσοστό σφάλματος της πολιτικής και τον αριθμό των επαναλήψεων

.



Εικόνα 6

3.2 Experiments

Η έρευνα διαχείρισης διαλόγου συνήθως αξιολογείται σε ένα μικρό σύνολο περιβαλλόντων. Ευτυχώς, ένα σύνολο εκτεταμένων προσομοιωμένων περιβαλλόντων διαχείρισης διαλόγου δημοσιεύεται στο ,το οποίο μπορεί να δοκιμάσει την ικανότητα των μοντέλων σε διαφορετικά περιβάλλοντα. Αυτά τα περιβάλλοντα υλοποιούνται σε μια εργαλειοθήκη ανοιχτού τομέα: PyDial.

Υπάρχουν τρεις διαφορετικοί τομείς: εργασίες αναζήτησης πληροφοριών για εστιατόρια στο Cambridge (CR) και στο San Francisco (SFR) και μια γενική εργασία αγορών για φορητούς υπολογιστές (LAP). Βασίζονται σε θέσεις υποδοχής, πράγμα που σημαίνει ότι η κατάσταση διαλόγου παραγοντοποιείται σε κουλοχέρηδες.

Η δεύτερη διαφορετική διάσταση της μεταβλητότητας είναι το ποσοστό σημασιολογικού σφάλματος (SER), το οποίο προσομοιώνει διαφορετικά επίπεδα θορύβου στο κανάλι εισόδου κατανόησης ομιλίας.

Οι μετρήσεις στην επόμενη ενότητα παρουσιάζουν το μέσο ποσοστό επιτυχίας για κάθε μοντέλο που εφαρμόζεται. Το ποσοστό επιτυχίας ορίζεται ως το ποσοστό των διαλόγων που ολοκληρώθηκαν με επιτυχία.

Αξιολογήσαμε τους αλγόριθμους μας στους ακόλουθους τρεις τομείς:

a) Cambridge Restaurants (CR), που είναι ο πιο κοινός τομέας στη βιβλιογραφία. Παράγει ένα διάνυσμα sumBS 268 διαστάσεων.
b) Εστιατόρια του Σαν Φρανσίσκο (SFR) που έχουν υψηλότερη διάσταση (636 διαστάσεων)
   sumBS διάνυσμα.
c) Φορητοί υπολογιστές (LAP) που είναι από τους πιο δύσκολους τομείς, ειδικά για υψηλότερο SER. Το LAP έχει ένα διάνυσμα sumBS 257 διαστάσεων.

3.2.1 Experiments on GP-Sarsa, DQN, and eNAC

Πίνακας 1:Success rates after 1000/4000 training dialogues. The results in bold are the best success rate.

after 1000 training dialogues				
		Success rate %		
SER		GP-Sarsa	DQN	eNAC
0%	CR	**98.0**	88.6	93.0
	SFR	**91.9**	48.0	85.8
	LAP	78.9	61.9	**84.2**
15%	CR	**91.4**	79.5	85.7
	SFR	**76.6**	42.4	73.6
	LAP	65.0	51.9	**71.0**
30%	CR	**84.9**	72.3	73.6
	SFR	**59.7**	35.6	55.2
	LAP	52.0	47.5	**56.3**
after 4000 training dialogues				
0%	CR	**99.4**	93.9	94.8
	SFR	**96.1**	65.0	94.0
	LAP	89.1	70.1	**91.4**

15%	CR	**95.1**	93.4	90.8
	SFR	81.6	60.9	**84.6**
	LAP	68.3	61.1	**76.6**
30%	CR	**89.6**	87.8	79.6
	SFR	64.2	47.2	**66.7**
	LAP	44.9	46.1	**64.6**

3.2.2 Experiments using sumBS/ AutoEncoders / Denoising AutoEncoders

SER	BS	CR	LAP	SFR
0%	sumBS	98.4%(±1.1)	86.8%(±3.8)	95.2%(±1.3)
	AE5	**99.3%(±0.5)**	**92.6%(±1.9)**	95.3%(±0.8)
	AE7	97.3%(±0.8)	90.9%(±1.7)	**95.4%(±2.1)**
15%	sumBS	96.4%(±2.2)	66.5% (±2.3)	81.6% (±1.6)
	AE5	**96.5%(±1.0)**	68.9%(±10.1)	89.3%(±1.8)
	DAE5	94.5%(±0.7)	**80.5%(±0.7)**	87.9%(±0.5)
	DAE7	91.9%(±2.9)	89.7%(±3.1)	**95.1%(±1.4)**
30%	sumBS	88.5%(±4.2)	51.4%(±9.3)	66.3%(±5.3)
	AE5	92.2%(±1.1)	50.1%(±10.4)	69.4%(±2.3)
	DAE5	92.1%(±0.7)	72.9%(±0.6)	84.0%(±0.8)
	DAE7	**92.9%(±2.4)**	**84.3%(±4.0)**	**94.9%(±1.28)**
45%	sumBS	78.0%(±3.4)	24.1%(±5.5)	53.9%(±6.8)
	AE5	78.1%(±3.3)	38.7%(±5.4)	36.9%(±7.9)
	DAE5	84.2%(±0.8)	76.3%(±0.8)	78.8%(±1.6)
	DAE7	**91.2%(±5.4)**	**88.0%(±2.9)**	**81.9%(±7.8)**

Πίνακας 2:Average dialogue success after 3000 dialogues using AE and DAE topologies. Standard deviation in parenthesis

3.3 Discussion

Όπως μπορούμε να δούμε τη δραστική απόκλιση των μέσων ποσοστών επιτυχίας στους τομείς LAP, CR και SFR για διαφορετικά επίπεδα SER και τις διαφορετικές τοπολογίες AE.

Από τον Πίνακα 1 μπορούμε να καταλάβουμε ότι περισσότεροι εκπαιδευτικοί διάλογοι ισοδυναμούν με υψηλότερα ποσοστά επιτυχίας. Μπορούμε να επιβεβαιώσουμε ότι οι μέθοδοι ενισχυτικής μάθησης χρειάζονται πολλή εκπαίδευση για να φτάσουν σε υψηλά ποσοστά και καθώς το ποσοστό ανεβαίνει το κόστος αυξάνεται εκθετικά. Τα αποτελέσματα δείχνουν ότι το eNAC έχει τις ισχυρότερες δυνατότητες γενίκευσης, έχοντας την καλύτερη απόδοση στα περισσότερα περιβάλλοντα πολλαπλών εργασιών. Τα μοντέλα που βασίζονται σε τιμές έχουν καλή απόδοση όταν εκπαιδεύονται με θορυβώδη δεδομένα και δοκιμάζονται σε καθαρά δεδομένα, με το DQN να προσεγγίζει την απόδοση πολύ κοντά στο eNAC. Ωστόσο, όταν εκπαιδεύεται σε καθαρά δεδομένα και δοκιμάζεται σε θορυβώδη δεδομένα, η απόδοση μειώνεται σημαντικά, ειδικά στους μεγαλύτερους τομείς. Αυτή η μείωση στην απόδοση είναι πιο σοβαρή για το GPSARSA. Υπάρχουν άλλοι αλγόριθμοι RL αλλά λόγω χρονικών περιορισμών χρησιμοποίησα τρεις βασικούς.

Στον Πίνακα 2 δοκιμάζουμε διαφορετικά Συστήματα πεποιθήσεων (BS) όπως το sumBS και μερικούς αυτόματους κωδικοποιητές με GPSARSA. Τα γραφήματα του Πίνακα 2 δείχνουν την εξέλιξη των μέσων ποσοστών επιτυχίας στους τομείς LAP11, CR και SFR για διαφορετικά επίπεδα SER και τις διαφορετικές τοπολογίες AE. Φαίνεται ότι σε όλα αυτά τα διαγράμματα, η πολιτική δείχνει μια σχετικά ομαλή και γρήγορη σύγκλιση για όλες τις αναπαραστάσεις. Μπορεί να φανεί ότι ο διαχειριστής διαλόγου επωφελείται από την αναπαράσταση vanilla AE5 για 0% SER, η οποία είναι συνεπής για οποιονδήποτε αριθμό επεισοδίων, αλλά δεν είναι σε θέση να παρέχει δυνατότητες ανθεκτικές στον θόρυβο για υψηλότερα SER. Από την άλλη πλευρά, καθώς η παρουσία θορύβου αυξάνει, η αναπαράσταση με βάση τις τοπολογίες DAE5 και DAE7 δείχνει μεγάλες δυνατότητες. Η απόδοσή τους είναι πολύ υψηλότερη, ιδιαίτερα στους δύσκολους τομείς του LAP11 και του SFR. Συγκεκριμένα, το DAE7 διατηρεί απόδοση κοντά στο 90% ακόμη και για επίπεδα θορύβου έως και 45%. Η κυριαρχία τους στην απόδοση είναι ακόμη πιο εμφανής στους δύσκολους τομείς των LAP11 και SFR, όπου η απόδοση του sumBS και της βανίλιας AE5 υποβαθμίζεται δραματικά.

5 ΒΙΒΛΙΟΓΡΑΦΙΑ

Dette, H. &. (1996, September ?). Wall and Siegmund Duality Relations for Birth and Death Chains with Reflecting Barrier. *Journal of Theoretical Probability*, σ. ??

Gaˇsi´c, M. (January 2011). *Statistical Dialogue Modelling.* Cambridge: Department of Engineering and Trinity Hall University of Cambridge.

Hamidreza Chinaei, B. C.-d. (2016). *Building Dialogue POMDPs from Expert Dialogues.* University of Toronto, Toronto, ON, Canada: Springer, Cham.

Jean-Baptiste, E. M. (June 2016). *Statistical Task Modeling of Activities of Daily Living for Rehabilitation.* Toronto: University of Toronto.

Stylianou, A. P. (2019). Single-model Multi-domain Dialogue Management with Deep Learning. *Advanced Social Interaction with Agents*, (σσ. 71-77).

Volodymyr Mnih, K. K. (2013). *Playing Atari with Deep Reinforcement Learning.* Cornell University.

Wieland Eckert, R. P. (June 1998). Using Markov decision process for learning dialogue strategies. *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98*, (σσ. 201–204).