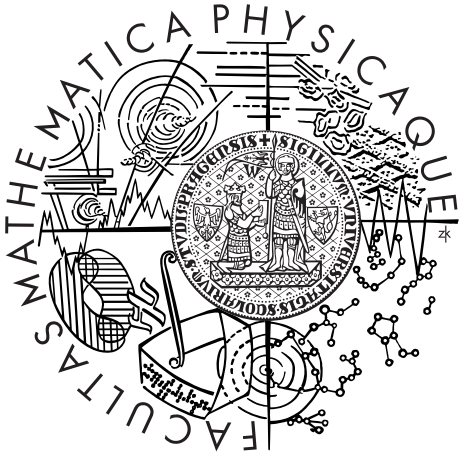
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE



David Marek

Implementace aproximativních bayesovských metod pro odhad stavu v dialogových systémech

Ústav formální a aplikované lingvistiky

Vedoucí diplomové práce:	Ing. Mgr. Filip Jurčíček, Ph.D.
Studijní program:	program
Studijní obor:	obor

Praha 2013

Poděkování.

Prohlašuji, že jsem tuto diplomovou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ..... Podpis autora

Název práce: Implementace aproximativních bayesovských metod pro odhad stavu v dialogových systémech

Autor: David Marek

Katedra: Ústav formální a aplikované lingvistiky

Vedoucí diplomové práce: Ing. Mgr. Filip Jurčíček, Ph.D., Ústav formální a aplikované lingvistiky

Abstrakt:

Klíčová slova:

Title:

Author: David Marek

Department: Institute of Formal and Applied Linguistics

Supervisor: Ing. Mgr. Filip Jurčíček, Ph.D., Institute of Formal and Applied Linguistics

Abstract:

Keywords:

Obsah

Úvod

Dialog je přirozený způsob dorozumívání a sdělování informací mezi lidmi. Počítač, který by dokázal vést dialog s uživatelem, byl vždy snem nejen příznivců vědecko-fantastické literatury. Už pro první počítače vnikaly programy, které se snažily využívat přirozenou řeč pro interakci s uživatelem. Jedním z takových programů byl například Eliza, program, který předstíral, že jej zajímá, co mu uživatel říká. Fungoval na principu rozpoznání textu pomocí gramatiky a následné transformace textu do promluv dle pravidel. Avšak gramatiky a pravidlové systémy se ukázaly nedostačné pro praktické aplikace a tak se vývoj přesunul do statistických metod. S využitím statistickým metod a metod strojového učení bylo možné začít s porozumíváním mluveného slova. Přijetí bylo zpočátku chladné a veřejnost byla

1. Teorie dialogových systémů

1.1 Dialogový systém

Dialogový systém je počítačový systém, který umožňuje uživatelům komunikovat s počítačem ve formě, která je přirozená a efektivní pro použití. Dialogové systémy stále mají spoustu problémů k překonání a pro praktické použití je třeba se uchýlit k několika předpokladům a zjednodušením. Prvním zjednodušením je doménová specializace, v současné době není možné vytvořit dialogový systém, který by se dokázal s uživatelem bavit o libovolném tématu. Vždy je potřeba při vývoji dialogového systému mít ontologii určující, jaké informace má systém poskytovat a o čem se může chtít uživatel bavit.

Další zjednodušení se týkají přímo dialogu. Předpokládá se, že dialog probíhá vždy mezi systémem a jedním uživatelem. Navíc se pravidelně střídají v obrátkách. Jedna obrátka dialogu je složená z jedné promluvy systému a jedné promluvy uživatele.

Příkladem dialogového systému může být systém pro nalezení spojení pomocí městské dopravy. Příkazu od uživatele může vypadat např. takto: „Chci jet z Malostranského náměstí na Anděl“. Dialogový systém se nyní může rozhodnout, zda-li mu zadané informace stačí pro nalezení spojení. V tomto případě systém stále neví, kdy chce uživatel jet. Může předpokládat, že uživatel už na zastávce stojí a tak tedy uživateli nalezne nejbližší spojení.

Důležitou vlastností dialogového systému je robustnost. Pokud budeme používat dialogový systém v přirozeném prostředí, musíme se vyrovnat s tím, že často nebude uživateli rozumět. Systém může informaci přeslechnout, anebo si nemusí být jistý tím, co slyšel. Dialogový systém se proto musí umět uživatele doptat na chybějící informace

1.2 Součásti dialogového systému

Dialogový systém se skládá z několika částí, které spolu komunikují. Na vstupu je zvukový záznam uživatele, o jeho převedení do textu se stará systém rozpoznávání řeči (ASR). Z textu je potřeba získat sémantické informace pomocí systému porozumění mluvené řeči (SLU). Nad sémanticky anotovanými informacemi už může pracovat dialogový manager (DM), který zvolí patřičnou odpověď. Výstupem dialogového manageru jsou informace, které se mají předat uživateli. O jejich převedení do textu se stará systém generování přirozené řeči (NLG). Do zvukového záznamu převede text syntetizér řeči (TTS).

1.2.1 Systém rozpoznávání řeči (ASR)

Systém rozpoznávání řeči slouží k převedení mluveného projevu do textové podoby. Po získání textové podoby je teprve možné se zabývat významem textu. Aktuálně nejlepší systémy jsou založené na pravděpodobnostním modelu a využívají skryté Markovské modely (HMM) k určení nejpravděpodobnější sekvence slov pro daný zvukový záznam. Pro tuto část dialogového systému existuje řada dostupných otevřených toolkitů, např. systém HTK [12], Kaldi [6] nebo SPHINX [9]. Existuje i celá řada komerčního software od firem jako IBM nebo Nuance.

Úspěšnost systému rozpoznání řeči je závislá na obtížnosti úlohy a na počtu trénovacích dat, pocházejících ze stejné domény. Pro obecnou doménu se problém stává mnohem těžší a je třeba velké množství dat. Word error rate (WER) je častá metrika pro počítání výkonu ASR. Pro spočítání WER je třeba nejprve provést zarovnání rozpoznaného a originálního textu. WER je pak počet slov, která jsou změněná, smazaná anebo přidaná, vydělený počtem slov v originálním textu. Systém Let's Go! [8] dosahuje průměrné WER 64.3%.

Systém rozpoznávání řeči může produkovat více hypotéz pro jeden vstup. Často existuje pro jeden zvukový záznam více možných slovních sekvencí, z kterých by mohl pocházet. Reprezentace možných hypotéz může být seznam slovních sekvencí s jejich věrohodností. Věrohodnosti jsou skóre přiřazené hypotézám, které určují jakou důvěru má systém rozpoznávání řeči ve správnost dané slovní sekvence. V ideálním případě je věrohodnost ekvivalentní aposteriorní pravděpodobnosti sekvence slov, dáno vstupní zvuk. Ovšem ne všechny rozpoznávače pracují na pravděpodobnostním principu a pak není možné od nich požadovat skutečné pravděpodobnosti.

Přidat příklad seznamu hypotéz

Další možností jak reprezentovat výstup je použití konfůzní sítě [1]. Konfůzní síť je vážený orientovaný graf, obsahující startovní a konečný vrchol a hrany označené slovy. Každá cesta ze startovního do konečného vrcholu vede přes všechny ostatní vrcholy. Váhy hran jsou pravděpodobnosti slova přiřazeného dané hraně. Hrany mohou obsahovat i prázdné slovo ϵ. Pravděpodobnost sekvence slov je součinem vah po cestě ze startovního do konečného uzlu. Výhodou konfůzní sítě je, že umožňuje v komprimované podobě uložit mnohem více hypotéz.

Přidat příklad konfůzní sítě

1.2.2 Porozumění mluvené řeči (SLU)

Jakmile má systém seznam možných hypotéz toho, co uživatel řekl, musí se pokusit porozumět, co tím uživatel myslel. Dialogový systém nepotřebuje vědět, co přesně uživatel řekl, důležité je pouze zjistit, co se uživatel snaží sdělit. Pokud například uživatel řekne "Chtěl bych nalézt spojení z Malostranského náměstí na Anděl", anebo "Jak se dostanu na Anděl ze zastávky Malostranské náměstí?", tak

výsledek je stejný, uživatel požaduje informace o spojení mezi dvěma zastávkami, i když v jednom případě jde o větu oznamovací a v druhém případě o otázku.

Sémantická reprezentace sdělení uživatele se nazývá dialogový akt (DA), skládá se z jedné nebo více položek dialogového aktu (DAI), které jsou spojené v konjunkci. Každá DAI se skládá z typu, názvu slotu a jeho hodnoty. Typy jsou doménově nezávislé, sloty a jejich hodnoty reprezentují koncepty ontologie. Příklad dialogového aktu z dialogového systému pro hledání restaurací:

```
hello()&inform(food="chinese").
```

Zde se dialogový akt skládá ze dvou položek, první položka má pouze typ *hello*, značící pozdrav. Druhá položka má typ *inform*, tzn. uživatel nás informuje o svém požadavku. Název slotu je *food* a hodnota je „chinese“, tedy uživatel nám říká, že hledá restauraci, kde servírují čínské jídlo.

Typů může být libovolné množství, ale existuje několik základních, jejichž použití je ustálené.

- *inform* — sdělujeme informaci, doplňujeme hodnotu do slotu,
- *request* — požadujeme od protějšku doplnění hodnoty pro dotazovaný slot,
- *confirm* — chceme potvrdit hodnotu slotu, potvrzení může být implicitní, anebo explicitní. Při explicitním potvrzení očekáváme odpověď buď „Ano“ nebo „Ne“, U implicitního, pokud se nám nedostane odpovědi předpokládáme, že protějšek souhlasí,
- *select* — žádáme protějšek, aby zvolil z nabízených možností.

Existuje široké množství technik, které lze použít pro porozumění mluvené řeči. Unifikace pomocí šablon anebo gramatiky jsou příklady ručně psaných metod. Metody založené na datech jsou například Hidden Vector State model [4], techniky strojového překladu [11], Combinatory Categorical Grammars [14] nebo Support Vector Machines [5].

1.2.3 Dialogový manager (DM)

Pokud už jsou pravděpodobné dialogové akty dekódovány, je třeba rozhodnout, co bude systém dělat. Komponenta tvořící tato rozhodnutí je dialogový manager. Odpověď systému je zakódována do formy dialogových aktů a nazývá se systémová akce.

Zvolená systémová akce je vybrána z množiny možných akci $a \in \mathcal{A}$ a závisí na vstupu, který systém obdržel z SLU. Tento vstup se nazývá pozorování $o \in \mathcal{O}$, protože obsahuje vše, co systém pozoroval o uživateli.

Zvolení správné akce potřebuje více znalostí než jen poslední pozorování. Celá historie dialogu a také kontext hrají důležitou roli. Dialogový manager bere na vše ohled pomocí udržovaní interní reprezentace celého pozorovaného dialogu. Tato reprezentace se nazývá dialogový stav, nebo také belief stav, značí se $b \in \mathcal{B}$. Aktuální dialogový stav závisí na přechodové funkci, která dialogový stav aktualizuje pro každé nové pozorování a systémovou akci. Přechodová funkce je tedy mapování $\mathcal{T} : \mathcal{B} \times \mathcal{A} \times \mathcal{O} \longrightarrow \mathcal{B}$. V této práci se budeme věnovat právě metodám aktualizace dialogového stavu.

Chování dialogového stavu definuje dialogová strategie π. Strategie určuje co má systém provést v závislosti na aktuálním dialogovém stavu. Obecně strategie vytvoří pravděpodobnostní rozložení přes možné akce. Pokud $\prod(\mathcal{A})$ značí množinu těchto distribucí, pak dialogová strategie bude zobrazení z dialogového stavu do této množiny, $\pi : \mathcal{B} \longrightarrow \prod(\mathcal{A})$.

Pozorování, dialogový stav a akce jsou číslovány podle obrátky. Pokud je časový okamžik důležitý, jsou pozorování, dialogový stav a akce z obrátky číslo t označeny o_t, b_t a a_t.

1.2.4 Generování přirozené řeči (NLG a TTS)

Posledním krokem v tahu dialogového systému je vytvoření odpovědi pro uživatele. Nejprve systém generování přirozené řeči převede dialogové akty na text. Následně je text převeden na zvuk pomocí textového syntetizéru řeči.

Nejjednodušším přístup ke generování přirozeného jazyka z dialogových aktů je použití šablon. Například pro dialogový akt `inform(type="x")` bude vytvořena šablona „Restaurace servíruje x jídlo“, kde „x“ bude nahrazeno například za „čínské“, „indické“, atd. Šablony se při generování osvědčily, protože počet možných dialogových je většinou zvládnutelný.

Při syntéze řeči existuje mnoho alternativ. Je možné použít segmenty řeči z databáze pro vygenerování zvuků tvořících dohromady celou sekvenci slov. Příkladem těchto systémů je Festival [2] nebo FLite [3].

Alternativní metodou syntézy je použití skrytých Markovských modelů pro generování zvuku, příkladem je HTS systém [13].

1.3 Dialogový manager

Struktura stavu je definována ontologií, ontologie popisuje doménu dialogového systému. Ontologie obsahuje

- koncepty, které se objevují v doméně,

- závislosti mezi jednotlivými koncepty,
- informace, na které se může systém dotázat,
- informace, o kterých může systém pouze informovat.

Závislosti mezi koncepty v doméně jsou důležité pokud může dojít ke sporu v požadavku uživatele. Příklad může být systém informací pro turisty, kde je možné hledat restaurace i hotely. Požadavek na počet hvězdiček je kompatibilní pouze s hledáním hotelu.

Dialogový stav tedy obsahuje informace závislé na ontologii, pro reprezentaci dialogu je důležité znát např.

- informace požadované uživatelem,
- informace potvrzené / zamítnuté uživatelem,
- o čem už systém informoval,
- co už systém potvrdil.

Výsledkem je tedy vytvoření rozšířeného stavu, který pro každý koncept obsahuje 3 podinformace, hodnoty o kterých bylo informováno, které byly potvrzeny, a které byly vyžádány. Také je potřeba sledovat další kontextové hodnoty jako např. pozdrav, poděkování, požadavek na více informací, anebo naopak na zopakování, atd.

V ideálním případě bychom měli mít informace o všem co bylo řečeno v dialogu, i kdy a v jakém pořadí.

DM se dělí na dvě části. První část se stará o udržování dialogového stavu, tzn. informace, které uživatel systému poskytl, historii dialogu, které informace už byly poskytnuty anebo např. potvrzeny či zamítnuty uživatelem. Dále obsahuje dialogovou strategii, která určuje příští akci v závislosti na dialogovém stavu.

Reprezentace stavu dialogu

Markovský rozhodovací proces (MDP) [7] je diskrétní, stochastický a kontrolovaný proces. V každém časovém okamžiku se systém nachází v nějakém stavu s. Uživatel provede akci a, dostupnou ve stavu s, a ta jej přesune náhodně do nového stavu s' a navíc dostane odměnu $r(s, s', a)$. Pravděpodobnost přechodu ze stavu s do stavu s' je dána přechodovou funkcí $p(s' \mid s, a)$.

Zobecněním MDP je částečně pozorovatelný Markovský rozhodovací proces (POMDP). Částečně pozorovatelný proto, že na rozdíl od MDP nevíme v jakém stavu se systém nachází. Naše představa o stavu je dána pouze pravděpodobnostním rozložením přes všechny možné stavy, tzv. belief $b(s)$. Výsledek provedení

akce tedy nezávisí pouze na přechodové pravděpodobnosti, ale také na belief stavu. Pro aktualizaci stavu je potřeba vysčítat přes všechny možné stavy

$$b(s') = \sum_{s} p(s' \mid s, a) b(s). \tag{1.1}$$

Problém dialogového systému je tedy POMDP [10], protože musíme zakomponovat naši neznalost cílů uživatele a nejistotu v rozpoznávání. Pro reprezentaci stavu dialogu existuje několik různých přístupů.

1.4 Grafické modely

2. Inference v grafických modelech

3. Učení parametrů

3.1 Grafický model

Máme vybraný faktor f, tento faktor je spojený s několika proměnnými $\boldsymbol{x} = (x_0, x_1, \ldots, x_{N_x})$ a množinami parametrů $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{N_\theta})$. Tento faktor reprezentuje podmíněnou pravděpodobnost:

$$f(\boldsymbol{x}, \Theta) = p(x_0 | x_1, \ldots, x_{N_x}; \Theta)$$

Rodičovské proměnné $x_1, \ldots, x_{N_x}$ označujeme jako $\boldsymbol{x'}$. Vektor $\boldsymbol{x'}$ určuje, která množina parametrů bude použita. Protože množiny parametrů jsou číslovány $1, \ldots, N_\theta$ a rodičovské proměnné $1, \ldots, N_x$, musí být pro vybrání správné množiny parametrů použito mapování $\rho(\boldsymbol{x'})$. Faktor pak může být zapsán zkráceně:

$$f(\boldsymbol{x}, \Theta) = p(x_0 | x_1, \ldots, x_{N_x}; \Theta) = \theta_{\rho(x'), x_0}$$

3.2 Výpočet marginálních pravděpodobností

Pro výpočet sdružené pravděpodobnosti používáme plně faktorizovanou distribuci. Pro každou proměnnou anebo množinu parametrů je její marginální pravděpodobnost rovna součinu zpráv přicházejících z faktorů, které jsou s danou proměnnou nebo množinu parametrů propojeny. Pro daný faktor je cavity distribuce $q^{\setminus}(x_i)$, popř. $q^{\setminus}(\boldsymbol{\theta}_i)$ rovna součinu zpráv ze všech ostatních faktorů do x_i, popř $\boldsymbol{\theta}_i$. Aproximovaná marginální pravděpodobnost proměnné je pak součinem cavity distribuce a zprávy z faktoru:

$$q(x_i) = q^{\setminus}(x_i) m_{f \to x_i}(x_i)$$

$$q(\boldsymbol{\theta}_i) = q^{\setminus}(\boldsymbol{\theta}_i) m_{f \to \theta}(\boldsymbol{\theta}_i)$$

Cavity distribuce je právě zpráva z proměnné, popř. množiny parametrů do faktoru.

$$m_{x_i \to f} = q^{\setminus}(x_i)$$

$$m_{\boldsymbol{\theta}_i \to f} = q^{\setminus}(\boldsymbol{\theta}_i)$$

3.2.1 Marginální pravděpodobnost proměnných

Pokud chceme aktualizovat hodnotu naší aproximace marginální pravděpodobnosti, tak je třeba minimalizovat její vzdálenost od skutečné marginální pravděpodobnosti:

$$p^*(\tilde{x}_j) = \sum_{\boldsymbol{x}:x_j=\tilde{x}_j} \int_{\boldsymbol{\Theta}} \prod_i q^{\setminus}(x_i) \prod_l q^{\setminus}(\boldsymbol{\theta}_l) f(\boldsymbol{x};\boldsymbol{\Theta}) \tag{3.1}$$

$$= \sum_{\boldsymbol{x}:x_j=\tilde{x}_j} \prod_i q^{\setminus}(x_i) \int_{\boldsymbol{\theta}_{\boldsymbol{\rho(x')}}} q^{\setminus}(\boldsymbol{\theta}_{\rho(\boldsymbol{x'})}) \theta_{\rho(\boldsymbol{x'}),x_0} \tag{3.2}$$

$$= \sum_{\boldsymbol{x}:x_j=\tilde{x}_j} \prod_i q^{\setminus}(x_i) \mathbb{E}_{q^{\setminus}}(\theta_{\rho(\boldsymbol{x'}),x_0}) \tag{3.3}$$

$$= \sum_{\boldsymbol{x}:x_j=\tilde{x}_j} \prod_i m_{x_i\to f}(x_i) \mathbb{E}_{q^{\setminus}}(\theta_{\rho(\boldsymbol{x'}),x_0}) \tag{3.4}$$

Rovnost (3.1) vychází z definice výpočtu marginální pravděpodobnosti ze sdružené pravděpodobnosti. V (3.2) byla použita definice faktoru, z integrálu byly vytaženy členy, které neobsahují Θ a nakonec bylo využito toho, že pro množiny parametrů, které nejsou spojeny s faktorem f, je jejich jejich cavity distribuce rovná marginální distribuci a tedy $\int_{\theta_i} q(\theta_i) = 1$. V (3.3) byla použita definice očekávané hodnoty.

Marginální pravděpodobnost proměnné x_i tedy je

$$p^*(\tilde{x}) = \sum_{\boldsymbol{x}:x_j=\tilde{x}_j} \prod_i m_{x_i\to f}(x_i) \mathbb{E}_{q^{\setminus}}(\theta_{\rho(\boldsymbol{x'}),x_0}) \tag{3.5}$$

Tady docházíme k výsledku, který je velmi podobný výpočtu marginální pravděpodobnosti v Loopy Belief Propagation algoritmu.

Zprávu z faktoru f do vrcholu x_j pak získáme vydělením zprávy z x_j z marginální pravděpodobnosti.

$$m_{f\to x_j}(x_j) = \sum_{\boldsymbol{x}:x_j=\tilde{x}_j} \prod_{i\neq j} m_{x_i\to f}(x_i) \mathbb{E}_{q^{\setminus}}(\theta_{\rho(\boldsymbol{x'}),x_0}) \tag{3.6}$$

3.2.2 Marginální pravděpodobnost parametrů

Pro množiny parametrů se jejich marginální pravděpodobnost spočítá podobně jako pro proměnné.

$$p^*(\tilde{\boldsymbol{\theta}}_j) = \sum_{\boldsymbol{x}} \int_{\boldsymbol{\Theta}:\boldsymbol{\theta}_j=\tilde{\boldsymbol{\theta}}_j} \prod_i q^{\setminus}(x_i) \prod_l q^{\setminus}(\boldsymbol{\theta}_l) f(\boldsymbol{x};\boldsymbol{\Theta}) \tag{3.7}$$

$$= \sum_{l\neq j} \sum_{\boldsymbol{x}:\rho(\boldsymbol{x}')=l} \prod_i q^{\setminus}(x_i) \int_{\boldsymbol{\Theta}:\boldsymbol{\theta}_j=\tilde{\boldsymbol{\theta}}_j} \prod_k q^{\setminus}(\boldsymbol{\theta}_k)\theta_{l,x_0} + \tag{3.8}$$

$$+ \sum_{\boldsymbol{x}:\rho(\boldsymbol{x}')=j} \prod_i q^{\setminus}(x_i) \int_{\boldsymbol{\Theta}:\boldsymbol{\theta}_j=\tilde{\boldsymbol{\theta}}_j} \prod_k q^{\setminus}(\boldsymbol{\theta}_k)\tilde{\theta}_{j,x_0}$$

$$= \left[\sum_{l\neq j} \sum_{\boldsymbol{x}:\rho(\boldsymbol{x}')=l} \prod_i q^{\setminus}(x_i)\mathbb{E}_{q^{\setminus}(\boldsymbol{\theta}_l)}(\theta_{l,x_0})\right] q^{\setminus}(\tilde{\boldsymbol{\theta}}_j) + \tag{3.9}$$

$$+ \sum_{\boldsymbol{x}:\rho(\boldsymbol{x}')=j} \prod_i q^{\setminus}(x_i)\tilde{\theta}_{j,x_0} q^{\setminus}(\tilde{\boldsymbol{\theta}}_j)$$

$$= w_0 q^{\setminus}(\tilde{\boldsymbol{\theta}}_j) + \sum_k w_k \tilde{\theta}_{j,k} q^{\setminus}(\tilde{\boldsymbol{\theta}}_j), \tag{3.10}$$

$$= w_0 m_{\tilde{\boldsymbol{\theta}}_j \to f}(\tilde{\boldsymbol{\theta}}_j) + \sum_k w_k \tilde{\theta}_{j,k} m_{\tilde{\boldsymbol{\theta}}_j \to f}(\tilde{\boldsymbol{\theta}}_j), \tag{3.11}$$

kde

$$w_0 = \sum_{l\neq j} \sum_{\boldsymbol{x}:\rho(\boldsymbol{x}')=l} \prod_i m_{x_i\to f}(x_i)\mathbb{E}_{q^{\setminus}(\boldsymbol{\theta}_l)}(\theta_{l,x_0}) \tag{3.12}$$

$$w_k = \sum_{\boldsymbol{x}:\rho(\boldsymbol{x}')=j,x_0=k} \prod_i m_{x_i\to f}(x_i) \tag{3.13}$$

Opět vycházíme z výpočtu marginální pravděpodobnosti ze sdružené pravděpodobnosti. V rovnici (3.8) jsme rozdělili sumu přes $\boldsymbol{x}$ na ty, pro které se ve faktoru použije množina parametrů $\tilde{\boldsymbol{\theta}}_j$ a na ty ostatní. Také jsme z integrálu vytknuli součin cavity distribucí pro proměnné. V dalším kroku (3.9) jsme opět použili toho, že integrál přes $\boldsymbol{\Theta}$ je ve skutečnosti několik integrálů přes jednotlivé množiny parametrů. A tedy je můžeme vložit mezi jednotlivé členy produktu cavity distribucí pro množiny parametrů. Ve výsledku získáme $q^{\setminus}(\tilde{\boldsymbol{\theta}}_j)\int_{\boldsymbol{\theta}_l} q^{\setminus}(\boldsymbol{\theta}_l)\theta_{l,x_0}$ a pak zbylé členy, které zmizí.

Docházíme k vyjádření skutečné marginální pravděpodobnosti, ve které není třeba integrovat přes všechny množiny parametrů, ale stačí jen očekávaná hodnota těchto parametrů.

3.3 Aproximace marginálních pravděpodobností

Stále tu ovšem zůstává problém, že spočítat aproximující distribuci $q(\boldsymbol{\theta}_j)$ může být příliš složité, protože skutečná marginální distribuce je směs několika distribucí a ta nemusí být v obecném případě vyjádřitelná. Je tedy třeba model dále aproximovat. Pro zjednodušení výpočtu jsou zprávy z faktoru do množiny parametrů, $m_{f\to\boldsymbol{\theta}_i}(\boldsymbol{\theta}_i)$, ve tvaru Dirichletovského rozdělení s parametry $\boldsymbol{\alpha}_{f\to\boldsymbol{\theta}_i}$:

$$m_{f\to\boldsymbol{\theta}_i}(\boldsymbol{\theta}_i) = Dir(\boldsymbol{\theta}_i;\boldsymbol{\alpha}_{f\to\boldsymbol{\theta}_i}) = \frac{\Gamma(\sum_j \boldsymbol{\alpha}_{f\to\boldsymbol{\theta}_i,j})}{\prod_j \Gamma(\boldsymbol{\alpha}_{f\to\boldsymbol{\theta}_i,j})} \prod_j \theta_{i,j}^{\boldsymbol{\alpha}_{f\to\boldsymbol{\theta}_i,j}-1} \tag{3.14}$$

kde Γ je Gamma funkce (zobecnění faktoriálu):

$$\Gamma(z) = \int_0^\infty t^{z-1}\exp(-t)\mathrm{d}t \tag{3.15}$$

Dirichletovské rozdělení bylo zvoleno, protože má důležité vlastnosti pro součin, které budou využity dále pro výpočet cavity distribuce a celkové aproximace. Pokud označíme aproximované faktory indexem β a každý bude mít vlastní parametry $\boldsymbol{\alpha}_{f_\beta\to\boldsymbol{\theta}_i}$, tak výsledná aproximace bude tvaru:

$$\begin{aligned} q(\boldsymbol{\theta}_i) &\propto \prod_\beta m_{f_\beta\to\boldsymbol{\theta}_i}(\boldsymbol{\theta}_i) && (3.16)\\ &\propto \prod_\beta \prod_j \theta_{i,j}^{\boldsymbol{\alpha}_{f_\beta\to\boldsymbol{\theta}_i,j}-1} && (3.17)\\ &\propto Dir(\boldsymbol{\theta}_i;\sum_\beta \boldsymbol{\alpha}_{f_\beta\to\boldsymbol{\theta}_i} - (|\beta|-1)\mathbf{1}) && (3.18)\\ &= Dir(\boldsymbol{\theta}_i;\boldsymbol{\alpha}_i) && (3.19) \end{aligned}$$

kde $\boldsymbol{\alpha}_i = \sum_\beta \boldsymbol{\alpha}_{f_\beta\to\boldsymbol{\theta}_i} - (|\beta|-1)\mathbf{1})$.

Při aktualizaci faktoru $\tilde{\beta}$ tedy cavity distribuce bude:

$$\begin{aligned} q^{\setminus\tilde{\beta}}(\boldsymbol{\theta_i}) &\propto \prod_{\beta\neq\tilde{\beta}} m_{f_\beta\to\boldsymbol{\theta}_i}(\boldsymbol{\theta}_i) && (3.20)\\ &\propto Dir(\boldsymbol{\theta}_i;\boldsymbol{\alpha}_i - \boldsymbol{\alpha}_{f_\beta\to\boldsymbol{\theta}_i} + \mathbf{1}) && (3.21) \end{aligned}$$

Naším cílem je nalézt parametry $\boldsymbol{\alpha}^*$ aproximované marginální pravděpodobnosti (3.19), které minimalizují vzdálenost od skutečné marginální pravděpodobnosti (3.10). Pro měření vzdálenosti mezi dvěma pravděpodobnostními rozloženími se používá Kullback-Leiblerova divergence:

$$KL(p\|q) = \int_{-\infty}^{\infty} p(x)\log\left(\frac{p(x)}{q(x)}\right)\mathrm{d}x \tag{3.22}$$

Pro nalezení minima použijeme algoritmus Expectation Propagation a budeme tedy minimalizovat $KL(p^*\|q)$.

Pokud se podíváme na skutečnou marginální pravděpodobnost $p^*(\boldsymbol{\theta}_i)$, zjistíme, že můžeme některé její členy upravit. Využijeme také vlastnosti gamma funkce $\Gamma(x) = (x-1)\Gamma(x-1)$.

$$w_j\theta_j Dir(\boldsymbol{\theta};\boldsymbol{\alpha}) \propto w_j\theta_j \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta_i^{\alpha_i - 1} \tag{3.23}$$

$$\propto w_j \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \theta_j^{\alpha_j} \prod_{i \neq j} \theta_i^{\alpha_i - 1} \tag{3.24}$$

$$\propto w_j \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \frac{\Gamma(\alpha_j + 1)\prod_{i\neq j}\Gamma(\alpha_i)}{\Gamma(1 + \sum_i \alpha_i)} Dir(\boldsymbol{\theta};\boldsymbol{\alpha} + \boldsymbol{\delta}_j) \tag{3.25}$$

$$\propto w_j \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \frac{\alpha_j\Gamma(\alpha_j)\prod_{i\neq j}\Gamma(\alpha_i)}{(\sum_i \alpha_i)\Gamma(\sum_i \alpha_i)} Dir(\boldsymbol{\theta};\boldsymbol{\alpha} + \boldsymbol{\delta}_j) \tag{3.26}$$

$$\propto w_j \frac{\alpha_j}{\sum_i \alpha_i} Dir(\boldsymbol{\theta};\boldsymbol{\alpha} + \boldsymbol{\delta}_j) \tag{3.27}$$

$$\tag{3.28}$$

Díky této úpravě lze p^* vyjádřit jako směs Dirichletovských rozdělení.

$$p^*(\boldsymbol{\theta}) = w_0^* Dir(\boldsymbol{\theta};\boldsymbol{\alpha}) + \sum_j w_j^* Dir(\boldsymbol{\theta};\boldsymbol{\alpha} + \boldsymbol{\delta}_j) \tag{3.29}$$

kde

$$w_0^* \propto w_0 \tag{3.30}$$

$$w_j^* \propto w_j \frac{\alpha_j}{\sum_i \alpha_i} \tag{3.31}$$

$$\sum_{i=0}^{k} w_i^* = 1 \tag{3.32}$$

Pro minimalizaci KL divergence mezi dvěma rozděleními z exponenciální rozdělení stačí, pokud se budou rovnat jejich postačující statistiky. Dokážeme jednoduše spočítat první dva momenty Dirichletovského rozdělení a tedy použijeme aproximaci a budeme počítat pouze s nimi a zbylé momenty zanedbáme. Je tedy třeba nalézt střední hodnotu a rozptyl proměnných z $p^*(\boldsymbol{\theta})$.

$$\mathbb{E}_{p^*}[\boldsymbol{\theta}] = \int \boldsymbol{\theta} p^*(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} \tag{3.33}$$

$$= \int \boldsymbol{\theta}(w_0^* Dir(\boldsymbol{\theta}; \boldsymbol{\alpha}) + \sum_j w_j^* Dir(\boldsymbol{\theta}; \boldsymbol{\alpha} + \boldsymbol{\delta}_j)) \, \mathrm{d}\boldsymbol{\theta} \tag{3.34}$$

$$= w_0^* \int \boldsymbol{\theta} Dir(\boldsymbol{\theta}; \boldsymbol{\alpha}) \, d\boldsymbol{\theta} + \sum_j w_j^* \int \boldsymbol{\theta} Dir(\boldsymbol{\theta}; \boldsymbol{\alpha} + \boldsymbol{\delta}_j) \, \mathrm{d}\boldsymbol{\theta} \tag{3.35}$$

$$= w_0^* \mathbb{E}_{Dir(\boldsymbol{\alpha})}[\boldsymbol{\theta}] + \sum_j w_j^* \mathbb{E}_{Dir(\boldsymbol{\alpha}+\boldsymbol{\delta}_j)}[\boldsymbol{\theta}] \tag{3.36}$$

Střední hodnotu proměnných $\boldsymbol{\theta}$ podle rozdělení p^* lze tedy spočítat jako vážený součet středních hodnot $\boldsymbol{\theta}$ podle jednotlivých Dirichletovských distribucí, z kterých se p^* skládá. Střední hodnota proměnné X_i podle Dirichletovského rozdělení je

První moment tedy máme spočítány, pro výpočet rozptylu můžeme využít přímo definici:

$$Var[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \tag{3.37}$$

Chybí nám tedy ještě výpočet střední hodnoty druhé mocniny proměnné $\boldsymbol{\theta}$ podle p^*. Můžeme ji vyjádřit z definice střední hodnoty.

$$\mathbb{E}_{p^*}[\boldsymbol{\theta}^2] = \int \boldsymbol{\theta}^2 p^*(\boldsymbol{\theta}) \, d\boldsymbol{\theta} \tag{3.38}$$

$$= w_0^* \int \boldsymbol{\theta}^2 Dir(\boldsymbol{\theta}; \boldsymbol{\alpha}) \, d\boldsymbol{\theta} + \sum_j w_j^* \int \boldsymbol{\theta}^2 Dir(\boldsymbol{\theta}; \boldsymbol{\alpha} + \boldsymbol{\delta}_j) \, d\boldsymbol{\theta} \tag{3.39}$$

$$= w_0^* \mathbb{E}_{Dir(\boldsymbol{\alpha})}[\boldsymbol{\theta}^2] + \sum_j w_j^* \mathbb{E}_{Dir(\boldsymbol{\alpha}+\boldsymbol{\delta}_j)}[\boldsymbol{\theta}^2] \tag{3.40}$$

Opět získáváme vážený součet středních hodnot podle Dirichletovských rozdělení. Střední hodnotu druhé mocniny proměnné podle Dirichletovského rozdělení lze opět jednoduše odvodit z definice.

$$\mathbb{E}_{Dir(\boldsymbol{\alpha})}[x_i^2] = \int x_i^2 Dir(\boldsymbol{x}; \boldsymbol{\alpha}) d\boldsymbol{x} \tag{3.41}$$

$$= \int x_i^2 \frac{\Gamma(\alpha_0)}{\prod_{j=1}^N \Gamma(\alpha_j)} \prod_{j=1}^N x_j^{\alpha_j - 1} d\boldsymbol{x} \tag{3.42}$$

Nyní jsme v podobné situaci jako v (3.23). Budeme postupovat stejně, vyjádříme nové Dirichletovské rozdělení.

$$\mathbb{E}_{Dir(\boldsymbol{\alpha})}[x_i^2] = \int \frac{\Gamma(\alpha_0+2)\alpha_i(\alpha_i+1)}{\alpha_0(\alpha_0+1)\Gamma(\alpha_i+2)\prod_{j\neq i}\Gamma(\alpha_j)} x_i^{\alpha_i+1}\prod_{j\neq i} x_j^{\alpha_j-1}\mathrm{d}\boldsymbol{x} \tag{3.43}$$

$$= \frac{\alpha_i(\alpha_i+1)}{\alpha_0(\alpha_0+1)}\int \frac{\Gamma(\beta_0)}{\prod_i\Gamma(\beta_i)}\prod_i x_i^{\beta_i-1}\mathrm{d}\boldsymbol{x} \tag{3.44}$$

$$= \frac{\alpha_i(\alpha_i+1)}{\alpha_0(\alpha_0+1)}\int Dir(\boldsymbol{x};\boldsymbol{\beta})\mathrm{d}\boldsymbol{x} \tag{3.45}$$

$$= \frac{\alpha_i(\alpha_i+1)}{\alpha_0(\alpha_0+1)} \tag{3.46}$$

Vyjádřili jsme $\Gamma(\alpha_0)$ a $\Gamma(\alpha_i)$ s pomocí $\Gamma(\alpha_0+2)$ a $\Gamma(\alpha_i+2)$

$$\Gamma(\alpha_0) = \frac{\Gamma(\alpha_0+2)}{\alpha_0(\alpha_0+1)} \tag{3.47}$$

$$\Gamma(\alpha_i) = \frac{\Gamma(\alpha_i+2)}{\alpha_i(\alpha_i+1)} \tag{3.48}$$

Následně jsme vytvořili nové parametry $\boldsymbol{\beta}$:

$$\beta_i = \alpha_i + 2 \tag{3.49}$$

$$\beta_{j\neq i} = \alpha_j \tag{3.50}$$

$$\beta_0 = \sum_i \beta_i \tag{3.51}$$

Nyní tedy dokážeme spočítat $\mathbb{E}_{p^*}[\boldsymbol{\theta}]$ a $\mathbb{E}_{p^*}[\boldsymbol{\theta}^2]$. Parametry aproximovaného rozdělení nalezneme následovně

$$\frac{\mathbb{E}[X_1]-\mathbb{E}[X_1^2]}{\mathbb{E}[X_1^2]-\mathbb{E}[X_1]^2} = \frac{\frac{\alpha_1}{\alpha_0}-\frac{\alpha_1(\alpha_1+1)}{\alpha_0(\alpha_0+1)}}{\frac{\alpha_1(\alpha_1+1)}{\alpha_0(\alpha_0+1)}-\frac{\alpha_1^2}{\alpha_0^2}} \tag{3.52}$$

$$= \frac{\frac{\alpha_1(\alpha_0+1)-\alpha_1(\alpha_1+1)}{\alpha_0(\alpha_0+1)}}{\frac{\alpha_0\alpha_1(\alpha_1+1)-\alpha_1^2(\alpha_0+1)}{\alpha_0^2(\alpha_0+1)}} \tag{3.53}$$

$$= \frac{\alpha_0\alpha_1(\alpha_0-\alpha_1)}{\alpha_1(\alpha_0\alpha_1+\alpha_0-\alpha_0\alpha_1-\alpha_1)} \tag{3.54}$$

$$= \alpha_0 \tag{3.55}$$

$$\alpha_i = \mathbb{E}[X_i]\alpha_0 \tag{3.56}$$

Z rovnice (3.52) vypočítáme sumu všech parametrů α_0. Protože střední hodnota proměnné z Dirichletovského rozdělení je právě $\frac{\alpha_i}{\alpha_0}$, tak jednotlivé parametry získáme z rovnice (3.56).

3.4 Algoritmus

Algoritmus 1 Expectation Propagation pro učení parametrů

Parametry zpráv z faktoru β do množiny parametrů $\boldsymbol{\theta}_i$ označíme $\boldsymbol{\alpha}_{f_\beta \to \boldsymbol{\theta}_i}$.
Parametry zpráv z množiny parametrů $\boldsymbol{\theta}_i$ do faktoru β označíme $\boldsymbol{\alpha}_{\boldsymbol{\theta}_i \to f_\beta}$.
Parametry marginální distribuce množiny parametrů $\boldsymbol{\theta}_i$ označíme $\boldsymbol{\alpha}_i$.
init
  Nastav zprávy mezi faktory a proměnnými na 1.
  Nastav parametry $\boldsymbol{\alpha}_{f_\beta \to \boldsymbol{\theta}_i}$ na 1.
  Nastav parametry $\boldsymbol{\alpha}_i$ na apriorní hodnotu.
end init
repeat
  Vyber faktor $f_{\tilde{\beta}}$, který se bude aktualizovat.
  Spočítej všechny zprávy z parametrů:
  for každý parametr $\boldsymbol{\theta}_i$ spojený s faktorem $f_{\tilde{\beta}}$ **do**
    Parametry zprávy z $\boldsymbol{\theta}_i$ do $f_{\tilde{\beta}}$: $\boldsymbol{\alpha}_{\boldsymbol{\theta}_i \to f_\beta} = \boldsymbol{\alpha}_i - \boldsymbol{\alpha}_{f_\beta \to \boldsymbol{\theta}_i} + 1$.
  end for
  Aktualizuj zprávy z faktoru do proměnných:
  for každou proměnnou X_i, spojenou s faktorem $f_{\tilde{\beta}}$ **do**
    Zpráva z $f_{\tilde{\beta}}$ do X_i podle (3.6):
      $\hat{f}(x_j) = \sum_{\boldsymbol{x}: x_j = \tilde{x}_j} \mathbb{E}_{q \setminus}(\theta_{\rho(\boldsymbol{x}'), x_0}) \prod_{i \neq j} m_{x_i \to f_{\tilde{\beta}}}(x_i)$
  end for
  Aktualizuj marginální pravděpodobnost parametrů:
  for každý parametr $\boldsymbol{\theta}_i$ spojený s faktorem $f_{\tilde{\beta}}$ **do**
    Spočítej parametry $\boldsymbol{\alpha}_i^*$ pro Dirichletovské rozdělení, které nejlépe aproximuje cílovou marginální distribuci (3.29). Metoda popsána v předchozí sekci.
    Parametry zprávy z $f_{\tilde{\beta}}$ do $\boldsymbol{\theta}_i$:
      $\boldsymbol{\alpha}_{f_\beta \to \boldsymbol{\theta}_i} = \boldsymbol{\alpha}_i^* - \boldsymbol{\alpha}_{\boldsymbol{\theta}_i \to f_\beta} + 1$
    Aktualizuj parametry marginální distribuce $q(\boldsymbol{\theta}_i)$
      $\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_i^* = \boldsymbol{\alpha}_{f_\beta \to \boldsymbol{\theta}_i} + \boldsymbol{\alpha}_{\boldsymbol{\theta}_i \to f_\beta} - 1$
  end for
  for každou proměnnou X_i, spojenou s faktorem $f_{\tilde{\beta}}$ **do**
    Aktualizuj zprávy z proměnných do faktoru:
      $m_{x_i \to f_{\tilde{\beta}}}(x_i) = \prod_{\beta \neq \tilde{\beta}} m_{f_\beta \to x_i}(x_i)$
  end for
until konvergence

Závěr

Literatura

[1] Bertoldi, N.; Federico, M.: A new decoder for spoken language translation based on confusion networks. In *Automatic Speech Recognition and Understanding, 2005 IEEE Workshop on*, IEEE, 2005, s. 86–91.

[2] Black, A.; Taylor, P.; Caley, R.; aj.: The Festival Speech Synthesis System, Version 1.4. 2. *Unpublished document available via http://www. cstr. ed. ac. uk/projects/festival. html*, 2001.

[3] Black, A. W.; Lenzo, K. A.: Flite: a small fast run-time synthesis engine. In *4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis*, 2001.

[4] He, Y.; Young, S.: Semantic processing using the hidden vector state model. *Computer speech & language*, ročník 19, č. 1, 2005: s. 85–106.

[5] Mairesse, F.; Gasic, M.; Jurcicek, F.; aj.: Spoken language understanding from unaligned data using discriminative classification models. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, IEEE, 2009, s. 4749–4752.

[6] Povey, D.; Ghoshal, A.; Boulianne, G.; aj.: The Kaldi Speech Recognition Toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, IEEE Signal Processing Society, Prosinec 2011, iEEE Catalog No.: CFP11SRW-USB.

[7] Puterman, M. L.: *Markov decision processes: discrete stochastic dynamic programming*, ročník 414. Wiley-Interscience, 2009.

[8] Raux, A.; Bohus, D.; Langner, B.; aj.: Doing research on a deployed spoken dialogue system: One year of Let's Go! experience. In *Proc. Interspeech*, 2006, s. 65–68.

[9] Walker, W.; Lamere, P.; Kwok, P.; aj.: Sphinx-4: A flexible open source framework for speech recognition. 2004.

[10] Williams, J. D.; Young, S.: Partially observable Markov decision processes for spoken dialog systems. *Computer Speech & Language*, ročník 21, č. 2, 2007: s. 393–422.

[11] Wong, Y. W.; Mooney, R.: Learning synchronous grammars for semantic parsing with lambda calculus. In *ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*, ročník 45, 2007, str. 960.

[12] Young, S.; Evermann, G.; Gales, M.; aj.: The HTK book. *Cambridge University Engineering Department*, ročník 3, 2002.

[13] Zen, H.; Nose, T.; Yamagishi, J.; aj.: The HMM-based speech synthesis system (HTS) version 2.0. In *Proc. of Sixth ISCA Workshop on Speech Synthesis*, 2007, s. 294–299.

[14] Zettlemoyer, L. S.; Collins, M.: Online learning of relaxed CCG grammars for parsing to logical form. In *In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL-2007*, Citeseer, 2007.

Seznam tabulek

Seznam použitých zkratek

Přílohy