MASARYKOVA UNIVERZITA V BRNĚ
FILOZOFICKÁ FAKULTA

PLIN021 Sémantická analýza v praxi

projekt

Vypracovala: Stanislava Sedláčková (118285)

semestr: jaro 2013

Obsah

1 Analýza dotazů....3

2 Podstata analýzy dotazů....4
2.1 Syntaktická analýza (*parsing*) a sémantická analýza....4
2.2 Přizpůsobení analýzy dotazů....5
2.3 Účelové relace (*special-purpose relations*)....6

3 Implementace pravidel v Prologu....7

4 Jádro dotazu a detekce LATu....8
4.1 Zlepšování detekce uvnitř dotazu....10
4.2 Extrakce LATů z kategorií....10
4.3 Spolehlivost odhadu LATu....11
4.4 Učící se LATy z předešlých dotazů v kategorii....11
4.5 Hodnocení....12

5 Klasifikace dotazů a detekce Qsekcí (*QSections*)....13
5.1 Klasifikace dotazů (*Question Classification*)....13
5.2 Metody pro zpracování speciální otázek (*Special Questions*)....17
5.2.1 Rozklad otázky a syntéza odpovědí....17
5.2.2 Omezení a slovní hříčky (*constraints and puns*)....18
5.2.3 Lexikální omezení....18
5.2.4 Omezující objekty (*constraint objects*)....18
5.2.5 Slovní hříčky....19
5.3 QSekce (*QSections*)....20
5.4 Hodnocení....21

6 Zdroje....23

Táto práce je založená na článcích IBM Výzkumného centra pro systém Watson zabývajícími analýzou dotazů systému Watson „Question analysis: How Watson reads a clue“ [1] a speciálními otázkami a metodami jejich analýzy: „Special questions and techniques [2]“. Práce může být rozsáhlejší kvůli uvádění konkrétních příkladů, nicméně bez jejich užití mi přišlo vykládání o metodách analýzy systému Watson nedostatečně srozumitelné. Z důvodu mé omezené znalosti odborné terminologie v dané oblasti raději uvádím při překladech v závorkách také originální anglické termíny. Stejně tak uvádím některé příkladové věty jak v češtině tak v angličtině, případně pouze v angličtině – nebylo-li v mých silách zaručit korektní překlad nebo kdy se jednalo o příklady příkazů v programovacím jazyce Prolog.

Odkazy na další články zabývajícími se systémem Watson, které jsou volně přístupné online, uvádím na konci práce, v kapitole Zdroje.

1 Analýza dotazů

U IBM Watson™ systému je aplikováno několik detekčních pravidel a klasifikátorů k rozeznání rozhodujících prvků/elementů dotazu:

1) část dotazu, která poukazuje na odpověď – jádro, centrum, zaměření (*focus*)
2) klasifikace dotazu do tříd (*QClasses*) podle typu (Q*uestion Classification*)
3) podmínky (kontext) dotazu, které naznačují, na jaký druh subjektu (kategorie) se dotazuje – lexikální typ odpovědi (*LAT, lexical answer types*)
1) a elementy dotazu, které hrají zvláštní roli a které mohou vyžadovat speciální zacházení, např.: vnořené subdotazy, které musí být zodpovězeny zvlášť (*Question Sections, QSections*)

Obecně tedy na začátku Watson přijme dotaz ve formě nestrukturovaného textu, který je poté zakódován jako strukturovaná informace, která se později využije jinými částmi Wastona. Téměř všechny části Watsona v určité míře závisí na informaci, kterou produkuje „analýza dotazu“ (*question analysis*). Analýza dotazů je postavena na základech univerzální syntaktické analýzy (*general-purpose parsing*) a komponent sémantické analýzy (*semantic analysis*). Obecně jsou tyto komponenty nezávislé na doméně, ale v některých případech byly upraveny na dotazy, které jsou specifické pro soutěž Jeopardy!

Jádro (*focus*), LAT (jako lexikální typ dotazu, *lexical answer types*), klasifikace dotazů do tříd (*Question Classification*, *QClasses*) a QSekce (*QSections*)[1] jsou nejdůležitější prvky analýzy dotazů. Jejich definice si lze lépe představit na příkladu (příkladová otázka ze hry Jeopardy!):

Kategorie: Básníci a poezie
Nápověda: Před tím než vydal sbírku „Songs of a Sourdough“ v roce 1907, pracoval jako bankovní úředník (1)

Jádro je ta část dotazu, která poukazuje na odpověď. V tomto případě je jádro dotazu „on“ (podst. jméno„he“)[2]. Jádro je například užito algoritmy, které se snaží sladit dotaz s potenciálně nosnou pasáží – pro správnou shodu by měla odpověď v dané pasáži ladit s jádrem dotazu.

LATy jsou podmínky dotazu (kontext), které označují na jaký druh subjektu se dotazuje (kategorie, typ odpovědi). Určující slovo (*headword*) jádra dotazu obecně určuje lexikální typ odpovědi (LAT), ale dotazy často obsahují i dodatečné zdroje LATů – v případě Jeopardy! domény jsou dodatečnými

1 Pro účely této práce volím vlastní kvazi-české překlady uvedených termínů s originálem v závorce
2 V případech kdy vícero slov odkazuje na odpověď je obecně postačující, pokud detekujeme pouze jeden fokus a poté použijeme univerzální koreferenční rozlišení (*general-purpose co-reference resolution*) pro nalezení dalších referencí

zdroji určujícími LAT kategorie dotazů (např. „Básníci a poezie“). Ve výše uvedeném příkladě (1) jsou LAT „on“, „úředník“ a „básník“. Watson používá LATy svými korekčními komponentami k určení toho, zda potenciální dotaz je instancí typu odpovědi.

Klasifikace dotazů (*Question Classification*) rozpozná, zda dotaz odpovídá jednomu nebo několika rozsáhlým typům. Uvedený dotaz z příkladu patří do jedné z nejrozsáhlejších tříd (*QClasses*), tzv. Faktoid (*Factoid*)[3]. Dalšími, hojně zastoupenými třídami v soutěži Jeopardy! jsou např.: Definice (*Definition*), Více možností (*Multiple-Choice*), Puzzle, Společné vazby (*Common Bonds*), Doplňte (*Fill-in-the-Blanks*) a Zkratky (*Abbreviation*). Tyto třídy dotazů (*QClasses*) se používají na vyladění procesu hledání odpovědi vyvoláním odlišných odpovědních technik, odlišných modelů strojového učení, nebo obojího. QSekce (*QSections*) jsou části dotazů, kterých interpretace vyžaduje zvláštní přístup. Některá z nejdůležitějších užití QSekcí je identifikace lexikálních omezení odpovědi (např. „4-letter“ nebo „3-word“) a rozklad dotazu na mnohonásobné subdotazy. Většina částí analýzy dotazů (*question analysis*) je implementována v Prologu. Tato implementace umožňuje analyzovat dotaz ve zlomku vteřiny, co je klíčové pro konkurenceschopnost v soutěži Jeopardy!

Téměř všechny systémy pro odpovídání na dotazy zahrnují nějaký podsystém pro analýzu dotazů. Velké množství těchto systémů bylo vyvinutých pod vlivem organizovaných hodnotících přístupů jako TREC/TAC (*Text Retrieval Conference/Text Analytics Conference*) a CLEF (*Cross-Language Evaluation Forum*). Tyto přístupy kladou důraz na dotazy typu Faktoid jako „Ve kterém filmu se poprvé objevil James Dean?“ nebo „Kolik obyvatel má Japonsko?“. Nicméně dotazy v soutěži Jeopardy! jsou mnohem komplexnější (složitější) a proto je přesná analýza dotazů mnohem obtížnější. Jeopardy! dotazy jsou navíc organizovány do kategorií a určit způsob, jakým má být název kategorie využitý (k určení významu), je náročný ale důležitý úkol zároveň a doposud se tomu žádné práce nevěnovaly.

2 Podstata analýzy dotazů

2.1 Syntaktická analýza (parsing) a sémantická analýza

Syntaktická a sémantická analýza systému Watson je složena z parseru gramatických kategorií (*Slot Grammar parser ESG – English Slot Grammar*[4]) s asociovaným rozhraním pro tvorbu struktur typu predikát-argument (*PAS – predicate-argument structure*), rozeznáváním pojmenovaných entit NER

3 Faktoid (*factoid*) je neověřená nebo nepřesná informace, která je (zejména v tisku) prezentována jako fakt, a která je tak časem akceptována jako pravdivá, zejména z důvodu častého výskytu, opakování. Definice např. zde http://www.thefreedictionary.com/factoid

4 Termín „slot grammar“ překládám jako „gramatické kategorie“ i když slovo „grammar“ zřejmě odpovídá překladu „gramém“ („základní lingvistický gramatický význam, který spolu s ostatními kategoriemi vytváří gramatické kategorie“ [3],[4]) . Dále se pak mluví o tzv. „slot- gramémech“ [5],[6] protože tyto gramémy jsou řazeny pomocí „slotů“ (gramatických relací) a pravidel pro jejich obsazování. V roce 2010 vydali v rámci IBM Research centra článek, který se zabýval využitím těchto slotů [5],[6]. V tomto článku se říká, že sloty mají dvě úrovně významu. Na jedné straně jsou sloty chápány jako syntaktické role frází ve větách; příklady takových slotů jsou: `subj -subjekt, obj -předmět přímý, iobj – předmět nepřímý, comp -rozvíjející člen predikátu, objeprep -předmět předložky, ndet – určitý/neurčitý člen`. Na druhé straně existují sloty (komplementární sloty; *complement slots)*, které mají sémantický význam (*significance*) a mohou plnit funkci určování pozice argumentů predikátů, které představují významy slov. Ku příkladu, mějme větu „Marie dala Honzovi knihu.“ Slovo „Marie“ obsadí slot `subj` (`subjekt`) slovesa „dala“, „Honzovi“ obsadí slot `iobj`, atd. Z tohoto pohledu představují sloty syntaktické role. Pro ilustraci sémantického významu si rozeberme význam slova „dát“, které lze v rámci predikátové logiky vyjádřit jako: *dát*(*e*, *x*, *y*, *z*) znamená „*e* je událost, kdy *x* dává *y* (předmět; něco) *z* (osobě; někomu)“. Logická reprezentace dané věty by byla: ∃*e*∃*y*(*kniha*(*y*) ^ *dát* (*e*, *Marie*, *y*, *Honza*)). Z toho pohledu mohou být sloty `subj`, `obj` a `iobj` považovány za názvy argumentů *x*, *y* a *z* slovesa *dát*. Lze ale také říct, že tyto sloty představují pozice argumentů pro predikáty slovesného významu. Déle viz [5],[6].

(*NER – named entity recognizer*), komponenty pro rozeznávání koreferencí a komponenty pro extrakci relací.

Pomocí ESG se každá věta parsuje (analyzuje syntakticky) do stromu, který znázorňuje jak povrchovou strukturu, tak vnořenou/vnitřní logickou strukturu. Každý uzel stromu má k sobě připojené:

1. slovo nebo víceslovný výraz (*multiword term*) s asociovaným predikátem nebo jeho logickým argumentem
2. seznam vlastností – některé jsou morfosyntaktické a některé sémantické
3. a pravý a levý modifikátor uzlu – každý se slotem, který zaplňuje

Vnitřní struktura spočívá v predikátech a jejich argumentech. Argumenty predikátů jsou další uzly stromu, ty ale mohou pocházet ze vzdálených pozic stromu nebo se může jednat o pasiv aktivní formy (logické formy) argumentů. Takové predikace jsou užitečné pro budování logických forem a práci se standardizovanými relacemi asociovanými s predikáty. Rozhraní pro tvorbu PAS zjednodušuje ESG analyzované úseky. V tomto rozhraní jsou „parsy“ (analyzované části vět) s malými obměnami v syntaxi namapované do společných forem, které pomáhají v následujícím hledání shod ve vzorech (*pattern matching*).

Ve vzorové větě (1) identifikují syntaktická a sémantická analýza (okrem jiného) predikace `vydání_sbírky(e1, on, ‘‘Songs of a Sourdough’’)` a `v(e2, e1, 1907)` [originál: `publish(e1, he, ‘‘Songs of a Sourdough’’)` a `in(e2, e1, 1907)`], kde druhá predikace říká, že událost vydání sbírky (*publish*) `e1` se odehrála v roce 1907. Dále také řeší i koreference dvou výskytů slov „on“ a „úředník“, identifikuje „Yukon“ jako geopolitickou entitu a „Songs of a Sourdough“ jako sbírku, a extrahuje relace jako `autorČeho(jádro, ‘‘Songs of a Sourdough’’)` a `časováSouvislost(vydání_sbírky(...), 1907)` [originál: `authorOf(focus, ‘‘Songs of a Sourdough’’)` a `temporalLink(publish(...), 1907)`].

2.2 Přizpůsobení analýzy dotazů

V soutěži Jeopardy! se všechny dotazy zobrazují velkými písmeny a přesně v takové podobě je i Watson přijímá. Pro člověka to nepředstavuje žádný problém, zatímco pro počítač je to docela výzva. Střídání velkých a malých písmen je zdrojem určitých informací, např. to pomáhá správně rozeznat názvy a jména osob. IBM do té doby využívala syntaktickou a sémantickou analýzu, které silně spoléhaly právě na střídání velkých a malých písmen. Tento problém byl řešen přidáním komponenty založené na statistickém odhadu střídání velkých a malých písmen, která byla testována na několika tisících příkladů. ESG byla dále v několika ohledech přizpůsobena tomu, jak jsou dotazy v soutěži Jeopardy! formulované. Například, Místo „wh“zájmen[5] se v otázkách Jeopardy! soutěže používají zájmena jako „tento/tato/toto/tito“ a „on/ona/to“, které vedou často k větným konstrukcím, které jsou v angličtině velmi vzácné: „Akciový kapitál, který je nízko rizikovou investicí je modrým „tímto“[6].“ Právě parser byl upravený tak, aby tento typ dotazů zvládal analyzovat. Dále se v Jeopardy! vyskytuje značné množství dotazů, které se skládají pouze z jmenných frází[7] (*noun phrase*). I když ESG parser byl vyvinut tak, aby zvládal segmenty textu s různými frázemi (ne jenom celé věty), otázky obsahující pouze jmenné fráze se vyskytují v

5 „Wh“ zájmena jsou anglická zájmena začínající na souhlásky wh- a patří sem: *what*, *which*, *where*, *when*, *who*, *whom*, *whose* (a někdy také: *whether*, *whatever*, *how* a *however*). Tyto zájmena jsou tázací nebo vztažné [8].

6 Neobratný překlad věty „A stock that is a low risk investment is a blue *this*“ - odpovědí je „Blue-Chip Stock“, tzv. modrý žeton, což je označení akcií zaběhnutých, největších a nejziskovějších společností obchodovaných na burze.

7 Jmenná fráze (*noun phrase*) je fráze, ve které vystupuje podstatné jméno nebo indefinitum jako určující člen celé fráze, nebo která má stejnou gramatickou funkci jako podstatné jméno [7].

otázkách soutěže Jeopardy! mnohem častěji než v průměrné angličtině. Parser byl tedy upravený tak, aby preferoval jmenné fráze za určitých podmínek. Například větu: „Počet básní Emily Dickinson, které povolila publikovat během svého života“ by předešlá verze ESG parseru považovala za slovesní frázi a k její analýze by přistupovala jako ve větě s hlavním slovesem „publikovat“. Samozřejmě, parser nebyl upraven tak aby byla opomenuta syntaktická analýza standardní angličtiny – tyto upravené funkce jsou zapnuty pouze v případě analýzy dotazů soutěže.

Dále bylo potřeba upravit také koreferenční komponentu, jelikož dotazy v soutěži často obsahují nevázaná zájmena jako indikátory jádra dotazu, která jsou pro normální text jinak nezvyklá. Jedná se o věty jako: „Oživil kosmonauta Dave Bowmana v jeho poslední novele 3001: The Final Odyssey“, kde zájmeno „jeho“ odkazuje na odpověď (Arthur C. Clarke) a ne na „Dave Bowman“. Toto bylo vyřešeno tím, že Watson nejdříve hledá jádro dotazu než přistoupí k řešení koreferencí a dále bylo také upraveno zpracování koreferencí s ohledem na jádro.

2.3 Účelové relace (*special-purpose relations*)

Komponenta pro extrakci relací byla také upravena ve dvou směrech pro správné zpracování výstředností dotazů v Jeopardy! soutěži. Jedna úprava je zaměřená na identifikaci a detekci účelových relací (*special-purpose relations*) se zvláštním významem pro Jeopardy! Na druhé straně, některé z primárních mechanizmů porovnávajících vzory, které se používají pro detekci relací, mohou být také použité pro identifikaci určitých (nerelačních; *non-relational*) aspektů dotazů soutěže, které jsou pak dále použity v klasifikaci dotazů (*Question Classification*) a zpracování speciálních dotazů. Účelové relace řeší specifické fráze v Jeopardy! soutěži, které vyjadřují vztahy docela běžně se vyskytujících v žánrech soutěžních dotazů. Jejich výskyt ale není vždy dostatečně častý na to, aby to odůvodnilo vývoj relačních extraktorů uvnitř syntaktické (parsing) a sémantické analýzy. Jejich frekvence v Jeopardy! je však dostatečně vysoká na to, aby jejich výskyt ovlivňoval následné komponenty. Takové relace (Table 1:Relace v dotazech v soutěži Jeopardy!) zahrnují následující: alternativní názvy pro elementy jádra, časovou aritmetiku v dotazech a geoprostorové relace. Dále byly detekovány další relace, nazývané `rdfTriple`, které identifikují jednotlivá fakta v rámci složených dotazů a které umožňují obecný přístup pro kontrolu pravdivosti těchto faktů.

Table 1: Relace v dotazech v soutěži Jeopardy!

Soutěžní otázka	Relace
Původně Flaviovský amfiteátr; stavba *tohoto* díla římské architektury začala za římského císaře Vespasiana kolem roku 72 n.l.	`altNázev(jádro, Flavovský amfiteátr)` [originál: `altName(focus, Flavian Amphitheatre)`]
Infarkt myokardu; více známý pod *tímto* jménem, je nejčastějším důvodem hospitalizace na jednotce intenzivní péče	`altNázev(jádro, infarkt myokardu)` [originál: `altName(focus, myocardial infarction)`]
V květnu 1898 oslavilo Portugalsko 400. výročí příchod *tohoto objevitele* do Indie	`výročíČeho(příchod tohoto objevitele do India, 400, květen 1898)` [originál: `anniversaryOf(this explorer's arrival in India, 400, May 1898)`]
Chile sdílí s *touto zemí* svou nejdelší hranici	`hraniceS(jádro, Chile)` [originál:

Soutěžní otázka	Relace
	`borderOf(focus, Chile)]`
USA koupila tuto skupinu ostrovů pojmenovanou podle ruského kapitána v roce 1867 a poté jej pronajímala společnosti pro lov tuleňů	`rdfTriple(koupit, USA, jádro)` [originál: `rdfTriple(buy, U.S., focus)`]

Určité třídy dotazů byly zpracovány samostatně pomocí sémantických rámců. Obecně řečeno, rámec shromáždí typické a charakteristické vlastnosti entit nebo událostí: příklady rámců, které byly použité při zodpovídaní dotazů v Jeopardy!, zahrnovaly knihy, prezidenty, země/státy a ceny (např.: „kdo byl oceněn/kdo vyhrál *co*, *za co* a *kdy*“). Myšlenka je taková, že když známe hodnoty a sémantiku některých rámcových slotů (*frame's slots*), uvažování specifické pro danou doménu může vést k hodnotě slotu prvku/elementu, který je jádrem dotazu. Přístup vnořených sémantických relací se nejlépe hodí pro detekci a pro vytváření instancí hodnot slotů. K tomuto účelu je součástí zodpovídání dotazů (*question-answering*) systém založený na sémantických rámcích. Ku příkladu otázka „Svoje 2 Oscary za herectví získal za roli drsného policistu v roce 1971 a surového šerifa v roce 1992“ vede k identifikaci následujících relací: `typOcenění (Oscar)`, `výherceOcenění (jádro)`, `oceněnáRole (drsný_policista)` a `kategorieOcenění (herectví)`[8]. Tyto relace jsou pak použity v mechanizmu vytváření instancí obecných rámců (v tomto případě by to vedlo ke dvěma instancím, které jsou identické až na slot `oceněnáRole`).

3 Implementace pravidel v Prologu

Tak jako jiné části ve Watsona i analýza dotazů je implementována jako samostatný systém seskupených komponent, který je postavený na Architektuře managementu nestrukturovaných informací UIMA (*UIMA – Unstructured Information Management Architecture*). Většina úkolů při analýze dotazů u Watsona je implementována jako pravidla nad rozhraním pro tvorbu struktur predikát-argument PAS (*PAS – Predicate-Argument Structure*) a různými externími databázemi jako je WordNet. Požadavek při volbě vhodného programovacího jazyka byl, aby bylo možné výhodně vyjádřit množství pravidel pro syntaktickou analýzu založenou na závislostech, včetně srovnávání vzdálených vztahů. Právě Prolog byl ideální volbou kvůli své jednoduchosti a možnosti vyjadřování.

V architektuře UIMA je obecné analýza CAS (*CAS – Common Analysis Structure*) dynamickou strukturou, která obsahuje nestrukturovaná data (tj. data, kterých zamýšlený význam ještě není odvozený) a strukturované informace odvozené z těchto dat kódovaných jako atributivní struktury (*feature structures*)[9]. Překlad obecné analýzy CAS do faktů v Prologu je přímá. Každá CAS atributivní struktura je přiřazena k unikátnímu celočíselnému ID. Každá vlastnost (*property*) této atributivní struktury se stane faktem ve formě `vlastnost_jméno(id, hodnota)` [origninál: `feature_name(id, value)`]. Pokud hodnota nějaké vlastnosti je zároveň jiná atributivní struktura, potom je ID cílové atributivní struktury použito jako hodnota faktu Prologu. Hodnoty pole (*array values*) jsou reprezentovány jako Prolog seznam. Každý nový fakt, který je výsledkem

8 originál: `awardType(Oscar), awardWinner(focus), awardRole(tough_cop), and awardCategory(acting).`

9 Atributivní struktura (*feature structure*) je v podstatě množina párů atribut-hodnota. Atributivní struktura může být reprezentována jako orientovaný acyklický graf s uzly, které korespondují s hodnotami proměnných a hranami korespondujícími s jejich názvy [9].

dotazování pravidel Prologu, je vrácen zpátky do CAS obecné analýzy jako nová atributivní struktura, která bude poté předána anotátorům v systému UIMA.

Na příklad PAS uzly, které vyprodukují anotátory syntaktické a sémantické analýzy, jsou reprezentovány jako fakta Prologu následovně (čísla reprezentují unikátní identifikátor PAS uzlu), věta (1), uvádím v originálu:

```
lemma (1, ‘‘he’’).          lemma (1, „on“).
partOfSpeech (1, pronoun).       částPromluvy (1, zájméno).
lemma (2, ‘‘publish’’).          lemma (1, „publikovat“).
partOfSpeech (2, verb).          částPromluvy (2, sloveso).
lemma (3, ‘‘Songs of a Sourdough’’). lemma (3, „Song of a Sourdough“).
partOfSpeech (3, noun).          částPromluvy (3, podstatné jméno).
subject (2, 1).             podmět (2, 1).
object (2, 3).           předmět (2, 3).
```

Některá pravidla byla nastavena tak, aby vyhledávala jádra dotazů, LATy a některé relace mezi elementy parsu (ve smyslu parse = analyzovaného úseku). Například zjednodušené pravidlo pro detekci relace `authorOf` lze v Prologu přepsat následovně:

```
authorOf(Author, Composition) :-
  createVerb(Verb),
  subject(Verb, Author),
  author(Author),
  object(Verb, Composition),
  composition(Composition) .-

  createVerb(Verb) :-
  partOfSpeech(Verb, verb),
  lemma(Verb, VerbLemma),
  [‘‘write’’, ‘‘publish’’, … ].
```

Predikáty `author (autor)` a `composition (sbírka)` vytváří omezení na uzly (tj. „on/he“, resp. „Song of Sourdough“) pro vyloučení uzlů, které nemohou být platně dosazeny do rolí `autor` a `sbírka` v relacích. Když aplikujeme toto pravidlo na příkladovou větu (1), dostaneme nový fakt `authorOf(1, 3)`, který je zaznamenán a dále je zpracován dalšími komponentami Watsona.

Použití Prologu značně vylepšilo produktivitu ve vývoji nových pravidel pro porovnávání vzorů (*pattern-matching*) a také umožnilo dosáhnout efektivitu zpracování, která je potřebná pro konkurenceschopnost ve hře Jeopardy! Bylo implementováno několik sad Prolog pravidel pro hloubkovou a povrchní extrakci relací, detekci jádra dotazu, detekci LATů, klasifikaci dotazů (*Question Classification*) a detekci QSekcí (*QSection*). Celkem bylo vytvořených více než 6000 klauzulí.

4 Jádro dotazu a detekce LATu

Implementace pro detekci základního jádra dotazu pozůstává z následujících vzorů – jádro dotazu je kurzívou a určující slovo (*headword*) je zvýrazněno tučně:

- Jmenná fráze s determinantem „tento/táto/toto“ (*this*) nebo „tito“ (*these*) určujícím postavení fráze v kontextu:
  DIVADLO: Nová hra na základě předlohy *této psí* ***klasiky*** *Sira Arthura Conana Doylea*

otevřela Londýnskou scénu v roce 2007.

- „tento/táto/toto“ (*this*) nebo „tito“ (*these*) jako zájmeno:
  80. LÉTA: V dubnu 1988 se stala společnost Northwest prvním leteckým přepravcem, která zakázala převážet ***toto*** na všech domácích linkách.
- Když je dotaz jmenná fráze, je celý dotaz označen jako jádro:
  AMERICKÁ LITERATURA: ***Počet*** *básní Emily Dickinson, které povolila publikovat během svého života*.
- Jedno ze zájmen „on/ona/jeho/jí/její“:
  NA ZÁPAD: (***Ona***) se připojila k Buffalo Bill Cody's Wild West Show poté, co ho potkala na Cotton Expo v New Orleans.
- Jedno ze zájmen „to/oni/jich/jej/jejich“:
  JÁ „PRVNÍ“!: Kongresovou ***to*** zakazuje vměšovat se do práva občanů na svobodu vyznání, projevu, shromažďovat se nebo na petici.
- Zájmeno „jeden“ (aplikovatelné zejména v angličtině, kde slovem one/ones občas nahrazujeme již dříve zmiňovaná podstatná jména):
  SLOVÁ NA 12 PÍSMEN: Leavenworth (město v USA), založené v roce 1895, je ***jedním*** takovým federálním. (originál: Leavenworth, established in 1895, is a federal ***one***.)

Pokud nelze aplikovat žádné z výše uvedených vzorů, dotaz nemusí mít jádro:

- NÁZVY FILMŮ – PÁRY: 1999: Jodie Foster a Chow Yun-Fat.

Navzdory tomu, že tato pravidla jsou přímá, tak požadavky z nich plynoucí pro parser mohou být náročné. Je kritické, aby parser správně přiřadil determinant „this“ k odpovídajícímu určujícímu slovu (*headword*) – například ve výrazu „*této psí* ***klasiky*** *Sira Arthura Conana Doylea*“ nenásleduje určující slovo ihned za „této“ (*this*). Je také velmi důležité, aby parser správně rozlišil dotaz s jmennou frází a slovesnou frází.

Detekce základního LATu obecně volí určující slovo jádra (*focus headword*) jako jediný LAT, s níže uvedenými výjimkami (jádro je kurzívou, LAT je tučně):

- Pokud je jádrem spojení slov/“konjunktů“ (*conjunct*), extrahujeme je:
  HENRY VIII: Henry nechal zničit hrobku v Canterburské kadedrále *tohoto* ***svatého*** *a* ***kanceláře*** *Henryho II*.
- “⟨Jádro⟩ X“. extrahuj LAT X pokud ⟨Jádro⟩ je některé z následujících třídy/jména/typu/druhu [originál: “⟨Focus⟩ for X“. extract LAT X when ⟨Focus⟩ is any of one/name/type/kind]: NÁ PIPI PIPI PIPI [originál:
  HERE, PIGGY, PIGGY, PIGGY]: Spousta matek přirovnala nepořádek v pokoji svého dítěte k *tomuto typu* ***příbytku*** *pro vepři.*
- “⟨Jádro⟩ X“. extrahuj LAT X pokud ⟨Jádro⟩ je některé z následujících jmen/slov/názvů [originál: “⟨Focus⟩ for X“. extract LAT X when ⟨Focus⟩ is any of one/name/type/kind]: PŮVOD NÁZVŮ SPOLEČNOSTÍ: James Church zvolil *toto jméno pro svůj* ***produkt***, protože symbol božského Vulkánu reprezentoval sílu.
- Pokud nebylo nalezeno žádné jádro a kategorií je jmenná fráze, vezmi určující slovo (*headword*) kategorie jako LAT:
  HEAVY METALOVÉ KAPELY: „Seek&Destroy“, „Nothing Else Matters“, „Enter Sandman“.

4.1 Zlepšování detekce uvnitř dotazu

Jednoduchá základní pravidla jsou rozumně přesná, ale samozřejmě ně stoprocentně neomylná, což se projevuje tím, že v textu je mnoho užitečných LATů, které nezachytí. Následují příklady, kdy základní vzory (*baseline patterns*) selhávají:

1) PÁRY: V dubnu 1997 se uskutečnila dražba věcí Clyde Barrowa pro financování stěhování jeho hrobu vedle ***jejího***.
2) OTEC ČAS[10] (400): 13. prosince 1961, se shledal Otec Čas s jeho *101 letou umělkyní* s příbuzným v ***její*** přezdívce.
3) ZLOČIN: Kapsář je zastaralé slovo pro *tento typ* ***zločince*** *pracujícího v davu*.
4) FTIPY: „Marmaduke“[11] je *toto plemeno* ***psa***.
5) PO OSTROVECH: I když jsou Indonésané povětšinou Muslimové, *toto* je převládající **náboženství** na Bali.
6) ZAPOMĚNÍ: Mýtické **řeky** Háda byly Styx a *tato jedna*, z které pili mrtví, aby zapomněli.
7) ZNÁMÍ AMERIČANÉ: Navzdory tomu, že nepřednesl žádnou volební řeč, (*on*) byl zvolen za **prezidenta** v roce 1868 s velkým náskokem.

Příklad 1) ilustruje obecný problém při detekci jádra a tím je výběr správného zájmena. Jednoduchá základní pravidla (*baseline rule*) vybírají lexikálně první, ale nesprávné jádro „jeho“. Pro výběr správného jádra se dále systém snaží stanovit, která zájmena jsou koreferenční s jmenovanými entitami v dotazu (např. „jeho“ k „Clyde Barrow“) a poté vybrat jako jádro zájmeno, které není vázáno na žádnou entitu. Když je jádro detekováno, je možné nasadit komponentu univerzálního anaforického rozlišování pro nalezení LATu ve formě zájmena, které by mohlo indikovat rod odpovědi, jako v příkladě 2).

Pokud jádro vyjadřuje podtřídu vztahu (příklady 3) a 4), jako nadřazená třída je označen LAT z vnořené předložkové fráze. LATy jsou často naznačeny z různých typů vztahů, ve kterých se nachází jádro, jako jsou koreference (příklad 5), množiny členství (příklad 6), nebo role jádra (příklad 7).

Další obtíže, se kterými se lze setkat při detekci LATu, je stanovení toho, zda má být LAT jedno slovo nebo víceslovný výraz (*multiword term*). Ve většině případů považujeme za LAT jedno slovo vyjma jakýchkoliv modifikátorů. Například, pokud by jádro dotazu byl výraz „tento americký prezident“, za LAT by bylo považováno slovo „prezident“. Modifikátory LATu by byly brány v potaz u některých dalších komponent, ale ne jako součást LATu samotného. Tento přístup ale může být v některých případech chybný (dejme tomu, že dotaz by obsahoval „tento vice prezident“ - v tom případě bychom nechtěli extrahovat LAT „prezident“). Z toho důvodu se považuje víceslovný výraz za LAT v případě, kdy nepředstavuje podtyp jeho určujícího slova (*headword*; jako ve „vice prezident“) nebo když modifikátor mění význam určujícího slova na vzácně se vyskytující význam (např. výraz „prime minister“ může být LATem, protože i když je podtypem významu „minister“, není často se vyskytujícím významem – platí pro Spojené Státy).

4.2 Extrakce LATů z kategorií

Jeden z nejobtížnějších úkolů je rozpoznat LATy, které se vyskytují v kategoriích. Jeopardy! kategorie někdy vyjadřují LATy (ale ne vždy) a není úplně lehké rozeznat, které slovo v kategorii může představovat LAT. Je několik málo vzorů, které přesně rozeznají LAT v kategoriích (např. „POJMENUJ X“ nebo „KDO JE X?“, nicméně tyto vzory pokrývají poměrně malé množství

10 Father Time: https://en.wikipedia.org/wiki/Father_Time
11 Marmaduke: https://en.wikipedia.org/wiki/Marmaduke

dotazů. V obecném případě lze říct, že slova z kategorií mohou být LATy, pokud splňují tři podmínky: odkazují na typ entit (např. název kategorie „Země“ víc než kategorie „Zeměpis“), typ entity je konzistentní s LATem dotazu (pokud dotaz LAT obsahuje) a poslední podmínkou je, že samotný dotaz se o daném typu nebo jeho instanci vůbec nezmiňuje.

Jako příklad lze uvést následující typy dotazů:

8) BRITŠTÍ **MONARCHOVÉ**: (*Ona*) přišla o velkou část vlasů než jí bylo 31 let.
9) GENERÁLNÍ PROKURÁTOŘI: Edmund Randolph pomohl navrhnout a ratifikovat Konstituci před tím, než se stal generálním prokurátorem *tohoto muže*.
10) PRVNÍ FILMY **HEREČEK**: Oklahoma!
11) AMERICKÁ **MĚSTA**: Je domovem Kentucké univerzity a dostihů Toyota Blue Grass Stakes.
12) AMERICKÁ MĚSTA: St. Petersburg je domovem Floridského každoročního turnaje v *této hře populární na palubách lodi*.

Kategorie v příkladu 8) obsahuje LAT „monarcha“, který vyjadřuje typ entity, který je kompatibilní s LATem „ona“. Příklad 9) má podobnou formu, ale „generální prokurátoři“ nejsou LAT. Naopak fakt, že kategorie je zmíněná v samotném dotazu, snižuje pravděpodobnost, že kategorie obsahuje LAT. Když se v příkladě 10) díváme pouze na název kategorie, zdá se, že slova „herečky“ a „film“ mají stejný potenciál být LATem. Pokud však můžeme zjistit, zda dotaz obsahuje slovo „film“ (nebo zda neobsahuje slovo „herečka“), můžeme říct, že slovo „herečka“ je LAT.

Příklady 11) a 12) mají stejnou kategorii, ale jenom kategorie příkladu 11) vyjadřuje LAT. V příkladu 12) není „město“ kompatibilní s LATem „hra“ a navíc přítomnost instance města v dotazu snižuje pravděpodobnost, že slovo „město“ bude LAT než téma.

4.3 Spolehlivost odhadu LATu

Pravidla rozpoznávání LATů – a zejména ty, které pracují s kategoriemi – mohou produkovat nesprávné závěry. Je to tím, že některá pravidla jsou více spolehlivá než jiná. Z tohoto důvodu je potřebné mít něco jako hodnotu spolehlivosti odhadu LATu, která může být použita jako váha v průběhu bodování (hodnocení) odpovědi. Pro tento účel byl „vytrénován“ logistický regresní klasifikátor, který používá manuálně anotované zlaté standardy LATů. Klasifikátor používá pravidla pro detekci jádra a LATů, které ohodnotil jako vlastnosti spolu s dalšími vlastnostmi z parsu, rozeznávání pojmenovaných entit NER (*NER – named entity recognizer*) a předešlé pravděpodobnosti toho, že vybrané slovo je LAT. LATy s nízkou spolehlivostí jsou filtrovány pro zlepšování přesnosti.

4.4 Učící se LATy z předešlých dotazů v kategorii

Watson může během soutěže přizpůsobit svou detekci LATů v kategorií učením se z předešlých dvojic (dotaz-odpověď), které byly odkryté v rámci (stejné) kategorie. Watson si pro každé slovo v kategorii (v názvu kategorie) vytvoří hypotézu, že by dané slovo mohlo být LATem a poté ověřuje, zda správná odpověď na předešlou otázku byla instancí tohoto typu. Pro Watson byly před-počítany statistiky z historických dotazů soutěže Jeopardy!, které vypovídají o pravděpodobnosti toho, že když slovo z kategorie bylo použito jako LAT v daném počtu předešlých dotazů, bude i nadále použito jako LAT v následujících dotazech. Tyto pravděpodobnosti byly spočteny odděleně pro případ, kdy pravidlo detekce LATu vybralo slovo z kategorie a pro případy, kdy toto pravidlo nebylo použito. Hodnota spolehlivost odhadu LATu slova z kategorie se rovná hodnotě odhadu této

pravděpodobnosti, která zároveň přepisuje hodnotu spočtenou logistickým regresním klasifikátorem. Jak již bylo zmíněno, LATy s nízkou spolehlivostí jsou filtrovány, co může produkovat zcela nové LATy, která nezaznamenala předešlá pravidla.
Například jedno tréninkové kolo, které Watson odehrál, obsahovalo kategorii „OSLAVY MĚSÍCE“ složenou z následujících otázek:

„Den D a Den Velké listiny práv a svobod“
„Národní[12] den filantropie a Dušičky“
„Národní den učitelů a Den Kentucky Derby“
„Den správních profesionálů a Národní den volna pro CPA (asi Certified Public Accountant)“
„Národní den kouzel a Nevadský den osla přijetí do unie“

Poté, co Watson viděl první otázky v kategorii, neinterpretoval kategorii jako vyjádření LATu a nesprávně detekoval „den“ jako LAT. Až poté co odpověděl a byl následně informován, že správná odpověď byla „červen“, zjistil, že odpověď zapadá do kategorie „měsíc“. To zvýšilo pravděpodobnost, že „měsíc“ bude LAT pro další nápovědy v kategorii, i když si pořád nemohl být jistý. Jak Watson „viděl“ další správné odpovědi v dané kategorii, tato pravděpodobnost rostla a Watson byl schopen vyhodnotit měsíce jako jeho nejlepší tip na odpověď pro další nápovědy v rámci kategorie.

4.5 Hodnocení

V IBM srovnali přesnost detekcí LATů (a tím pádem nepřímo i detekcí jádra, na kterém závisí) mezi základními pravidly pro detekci jádra a LATu (popsaných v předešlém textu) a celkového Watson systému. Pro vyhodnocení byla použita sada 9128 manuálně anotovaných otázek. Pro učení a vyhodnocování statistického LAT klasifikátoru bylo použito desetinásobné křížové validace (*ten-fold cross validation* nebo *k-fold cross validation*). Výstupy jsou znázorněné v Table 2: Vyhodnocování účinnosti detekce LATu a metriky byly definovány následovně:

$$Přesnost = \frac{počet\ správně\ detekovaných\ LATů}{počet\ detekovaných\ LATů} \tag{2}$$

$$Porkytí = \frac{počet\ správně\ detekovaných\ LATů}{počet\ LATů\ v\ manuálně\ anotovaném\ souboru} \tag{3}$$

$$F_1 = \frac{2(Přesnost)(Pokrytí)}{Přesnost + Pokrytí} \tag{4}$$

$$Pokrytí/otázka = \frac{počet\ otázek\ s\ alespoň\ jedním\ správně\ detekovaným\ LAT}{počet\ otázek\ s\ alespoň\ jedním\ manuálně\ anotovaným\ LAT} \tag{5}$$

Výsledky ukazují, že základní vzorce jsou dostatečně přesné ale jejich pokrytí významně zaostává. Právě na oblast zvyšování pokrytí (se zachováním úrovně přesnosti) byly zacílené snahy pro

12 U nás by se zřejmě hodil více termín „státní“, nicméně pro územní platnost různých významných dnů pro celé USA je snad více odpovídající termín „národní“.

vylepšení detekce jádra a LATu v systému Watson. Vylepšení pokrytí na otázku (*recall per question*) ukazuje, že u 6,5% dotazů typu Jeopardy! (tedy dotazů jejichž syntax je specifická pro tuto soutěž a často se liší od přirozeného jazyka) detekoval Watson správně LAT v případech, která nezachytila základní pravidla (*baseline rules*). Očekává se, že právě toto vylepšení zvýší celkové šance Watsona správně odpovědět na dotazy.

Table 2: Vyhodnocování účinnosti detekce LATu

	Základní pravidla	**Watson**
Přesnost (*Precision*)	0,817	0,829
Pokrytí (*Recall*)[13]	0,613	0,766
F_1	0,700	0,796
Pokrytí na otázku [*per question*]	0,840	0,905

Další z častých chyb je případ, kdy lidský anotátor určí za LAT víceslovný výraz (*multiword term*), zatímco Watson jako LAT určí jedno slovo (*single word*). Toto rozhodnutí může být subjektivní podle toho, jak se nadefinuje LAT. Jako příklad můžeme uvést rozpor v případě „zpěvák“ a „hlavní zpěvák“ (*lead singer*) nebo „orgán“ a „zákonodárný orgán“. Pokud zvolíme volnější způsob hodnocení a to takový, kdy LAT je ohodnocen jako správný, pokud se shoduje s jedním ze slov víceslovného zlatého standardu LAT, potom stoupne hodnota F_1 o 0,02 u Watsona a o 0,01 u základních pravidel (*baseline rules*).

5 Klasifikace dotazů a detekce QSekcí

V soutěži Jeopardy! se často vyskytují otázky, které není vhodné analyzovat jako celek, ale je naopak výhodné, analyzovat některé jejich části samostatně. Právě klasifikace dotazů (*Question Classification*), jejich rozklíčování do tzv. QTříd (*QClasses*) a identifikace QSekcí (*QSections*) je jednou z důležitých částí celé analýzy dotazů Watsona.

5.1 Klasifikace dotazů (*Question Classification*)

IBM definovala pro analýzu dotazů množinu QTříd (*QClasses*) – co jsou třídy dotazů. V některých případech je například nemožné odpovědět na otázky s pomocí standardního faktoidního přístupu (viz str.4, poznámka pod čarou č. 3) a identifikace těchto speciálních typů dotazů je nezbytná, např.:

13) PŘED A PO: Benátský cestovatel z 13. století, který je zároveň názvem vrchní části oblečení (*top*) s krátkým rukávem a límečkem od Ralpha Laurena.
14) NEJJIŽNĚJŠÍ HLAVNÍ MĚSTO: Helsinki, Moskva, Bukurešť

Správná odpověď otázky 13) („tričko Marco Polo“) se ve zdrojích Watsona ani neobjeví, tudíž odpovědět na tuto otázku standardním faktoidním přístupem nemůže nikdy uspět. Standardní

13 V terminologií IBM je používán anglický termín „recall“. Podle wiki [10], se termíny *precision* a *recall* používají v oblasti strojového učení. Jako *precision* se zde rozumí zlomek obdržených instancí z těch, které jsou relevantní a jako *recall* se označuje zlomek relevantních instancí ze všech, které jsou obdržené. *Recall* se někdy označuje i jako *sensitivity* (citlivost) a proto jsem se rozhodla právě pro tento překlad. Více o pojmech *recall* a citlivost (*sensitivity*) na [10] a [11]. Poznámka: „citlivost“ je po konzultaci s RNDr. Zuzanou Neveřilovou nahrazena „pokrytím“

faktoidní systém vždy selže i u otázky 14), protože správná odpověď („Bukurešť“) se objeví přímo v otázce, co je ve většině otázek primárně vyloučeno.

V soutěži se dále vyskytují třídy dotazů, ke kterým lze přistupovat standardním faktoidním způsobem, ale parciální analýza dotazů může být pro ně vhodnější. V Table 3:QTřídy (QClasses) je výčet QTříd Watsona spolu s jejich frekvencí výskytu, které byly naměřené manuálně na vzorku asi 3500 Jeopardy! otázek.

Table 3: QTřídy (QClasses)

QTřída (QClass)	**Popis**	**Příkladová otázka (správná odpověď)**	**Frekvence výskytu (%)**
DEFINICE (*Definition*)	Otázka, která obsahuje definici odpovědi	STAVEBNICTVÍ: Používá se při výstavbě cest nebo jako povrchová vrstva střech pro jejich voděodolnost. („asfalt“) STAVEBNICTVÍ: Název tohoto předmětu, který podepírá trámy, doslova znamená „něco, co nese“. („nosník“)	14,2
KATEGORIE – RELACE (*Category – relation*)	Odpověď má sémantický vztah (relaci) k otázce, kde tento vztah (relace) je specifikovaný v kategorii	BÝVALÍ GUVERNÉŘI STÁTŮ: Nelson A. Rockefeller. („New York“) ZEMĚ PODLE NOVIN: Haaretz, Yedioth Ahronoth (Izraelská periodika). („Izrael“)	7,2
FITB	Otázky typu Doplňte (*Fill-in-the-blank* – FITB) vyžadují doplnění nějaké fráze	DOPLŇTE: Autorství přisuzováno Lincolnovi: „____ je silnější než kulka.“ („volební lístek“) SHAKESPEARE A LÁSKA: „Ne, že bych Caesara miloval méně“, řekl Brutus, „ale 'tohle město' miluji víc.“ („Řím“)	3,8
ZKRATKY (*Abbreviation*)	Odpověď je význam zkratky v otázce	VOJENSKÉ ZÁLEŽITOSTI: Zkratkou SAS se označuje elitní britská vojenská jednotka podobná americké Delta Force. („Speciál Aire Servise“)	2,9
PUZZLE ~HÁDANKY (*Puzzle*)	Otázka typu puzzle – odpověď vyžaduje odvození, syntézu, metaforické a/nebo metonymické uvažování, atd.	PŘED A PO: Benátský cestovatel z 13. století, který je zároveň názvem vrchní části oblečení (*top*) s krátkým rukávem a límečkem od Ralpha Laurena. („tričko Marco Polo“) SCRABBLE SLOVA S NEJVYŠŠÍM BODOVÁNÍM: *Zoom*, *quiz* nebo *heaven*. („quiz“ - pro korektní odpověď bylo nutné dodržet originální znění.)	2,3

QTřída (QClass)	**Popis**	**Příkladová otázka (správná odpověď)**	**Frekvence výskytu (%)**
ETYMOLOGIE (*Etymology*)	Otázky, které se ptají na anglická slova odvozená z cizího slova mající daný význam	JSTE FOOD“E“? Ze španělského slova pro „péct v pečivu“ jde o jihoamerický ekvivalent calzone (plněná pizza, pizza kapsa). („an empada“)	1,9
SLOVESA (*Verb*)	Otázka tázající se na sloveso	NE TAK UPLNĚ SMRTELNÝ HŘÍCH: Napsat celý text e-mailu velkými písmeny je prohřešek internetové etikety, který naznačuje, že daná osoba provádí tuto činnost („křičení“ - v češtině je právě vhodnější překlad „křik“~provozuje křik)	1,5
PŘEKLAD (*Translation*)	Otázka vyžadující překlad slova nebo fráze z jednoho jazyka do jiného	OVOCE VE FRANCOUZŠTINĚ: Pomme. („jablko“)	1,1
ČÍSLA (*Number*)	Odpovědí je číslo	PŘEVODY JEDNOTEK: Jedna osmina kruhu je právě tolik stupňů. („45“)	1,0
VZÁJEMNÉ VZTAHY (*Bond*)	Dotaz je zaměřený na společné vlastnosti skupin entit	JEDLÉ SPOLEČNÉ VLASTNOSTI: Mungo, lusk, řetězec. („fazole, bob“, „*bean*“)	0,7
VÍCE MOŽNOSTÍ	Samotná otázka obsahuje několik možných odpovědí, z který je jen jedna správná	NEJJIŽNĚJŠÍ HLAVNÍ MĚSTO: Helsinki, Moskva, Bukurešť. („Bukurešť“) OSCAR, GRAMMY NEBO OBOJÍ: Mickey Rooney. („Oscar“)	0,5
DATA	Otázka se ptá na datum nebo rok	-NÁCTÁ LÉTA (*The Teens – '13-'19*): První světová válka skončila v listopadu tohoto roku. („1918“)	0,3

U těchto QTříd je využito několik technik. Rozpoznávače (*recognizers*) jsou vzájemně nezávislé a tím více než jedna QTřída může být přidružena ke kterékoliv otázce. Existují určité párové nekompatibility, které jsou vynucené konsolidačním procesem QTříd (*QClassConsolidator process*), který běží po rozpoznání a který dále odstraní méně preferované páry ze vzájemně nekonzistentních párů QTříd.

QTřídy „PUZZLE“, „VZÁJEMNÉ VZTAHY“, „FITB“ a „VÍCE MOŽNOSTÍ“ jsou docela standardně zastoupené v Jeopardy! soutěži a proto jsou primárně detekovány pomocí regulárních výrazů. Možná právě proto, že i lidé sami potřebují jakýsi signál ke změně techniky pro hledání odpovědi, otázky typu „PUZZLE“ a „VZÁJEMNÉ VZTAHY! jsou téměř vždy uvedené známými frázemi (např. „PŘED A PO“, „PŘESMYČKY“ nebo „SPOLEČNÉ VLASTNOSTI“) pravděpodobně i s modifikátorem určujícím doménu v samotném názvu kategorie. Dále jsou

aplikovány vzorce pro jednoduché regulérní výrazy (*simple regular expression patterns*), které detekují tyto jak výrazy, tak pozorovaná a očekávaná synonyma (v originále: „JUMBLED“ a „SCRUMBLED“). „FITB“ („doplňovačky“) se detekují pomocí regulérních výrazů nad textem otázky, které nejčastější odpovídají citovanému výrazu přilehlého jádru. Otázky typu „VÍCE MOŽNOSTÍ“ detekuje systém buď když název kategorie udává sekvenci (obvykle tří) položek a otázka naznačuje vztah pro výběr správné odpovědi, nebo opačně.

Další QTřídy, které nemají standardní zastoupení, jsou rozpoznávané pomocí syntaktických pravidel odpovídajícím struktuře predikát-argument (*PAS, Predicate-Argument-Structure*). QTřída „ETYMOLOGIE“ je zachycena výskytem vzorce „Z ⟨jazyka⟩ pro ⟨výraz⟩, ⟨jádro⟩ …, “ a dalších podobných vzorců. QTřídy „SLOVESA“ jsou rozeznána v případech, kdy je jádro odpovědi předmětem slovesa „dělat“ (tak jako v příkladu v Table 3:QTřídy (QClasses)) nebo když se jedná o otázku ve tvaru definice obsahující infinitiv (např.: „Toto slovo na 7 písmen znamená předpovídat nebo tušit“). Výskyt QTřídy „PŘEKLAD“ může být detekován buď vzorcem nad otázkou typu PAS („⟨jádro⟩ is ⟨jazyk⟩ pro ⟨výraz⟩“), nebo v případem kdy je otázka pouze jednoduchá fráze bez jádra a kategorie specifikuje jazyk, do nebo z kterého má být překládáno.
QTřídy „ČÍSLA“ a „DATA“ závisejí pouze na LAT a jsou detekovány, když LAT je položkou manuálně vytvořeného seznamu LATů, které mají čísla a data jako instance.
QTřídu „DEFINICE“ je samo o sobě těžké identifikovat s velkou přesností. Nicméně toto je kompenzované tím, že přístup při odpovídání na standardní dotazy je rozumně úspěšný při zpracování těchto dotazů a tedy přesnost v jejich detekci není tak kritická jako je tomu u jiných QTříd. Otázky typu „DEFINICE“ jsou identifikované pomocí klíčových frází jako „je výrazem pro“ a „znamená“ nebo pomocí několika málo opakujících se kategorií typu „NÁPOVĚDY V KŘÍŽOVKÁCH“.
QTřída typu „KATEGORIE-RELACE“ je zaznamenána v případě, že text otázky je jedna entita nebo krátký seznam entit (bez jádra). Tento typ QTříd je potlačený pokud je aplikovaná specifičtější QTřída.

Zvláštní pozornost je věnována QTřídám typu „ZKRATKY“. Posuďme případy:

15) ŽVÝKÁNÍ „TUKU“ (Originál: CHEWING THE „FAT“): Zkratka CFS skrývá diagnózu, která je také nazývána „zloděj vitality“. (Odpověď: CFS – *Chronic Fatigue Syndrome*, chronický únavový syndrom).
16) ZKRATKY: Na náhrobku: RIP. (Odpověď: *Rest In Peace* – Odpočívej v pokoji)
17) YANKEE MAGAZÍN: Článek s názvem „Smekání klobouku před Danbury“ v tomto stavu (státě) říká, jak JFK pomohl zničit průmysl jeho bezhlavými[14] způsoby.

Watson označí příklady 15) a 16) za QTřídy typu „ZKRATKY“ a jako zkratky, kterých význam je odpovědí identifikuje „CFS“, resp. „RIP“. V otázce 17) se také vyskytuje zkratka JFK, ale tato otázka není označena typem QTřídy „ZKRATKY“. Schopnost rozlišit mezi těmito příklady je klíčová, protože význam zkratky může být na jedné straně správná odpověď (otázky 15) a 16)), na druhé straně to může představovat velmi hloupou odpověď (otázka 17)).

Rozpoznání QTřídy „ZKRATKY“ je ve své podstatě obdobná způsobu detekce LATů: jedná se o kombinaci rozpoznávání na základě pravidel a statistického klasifikátoru (*statistical classifier*), který provede konečné rozhodnutí v závislosti primárně na naučené spolehlivosti každého pravidla. Rozpoznávání na základě pravidel obsahuje vzorce pro regulérní výrazy, které zachycují kanonické způsoby vyjádření otázek se zkratkami nebo pravidla pro shodu s PAS otázek k zachycení složitějších syntaktických variací v otázkách zaměřených na význam zkratek.

14 Originál: *bareheaded* – dle slovníku se jedná o stav bez pokrývky hlavy; může jít o nějaký ustálený výraz v AJ

5.2 Metody pro zpracování speciální otázek (*Special Questions*)

Procedury, které se používají pro zpracování speciálních otázek u Watsona mohou být rozděleny do dvou rozsáhlých kategorií: metody, které dělí typy otázek ve dvojí (metoda rozkladu otázky a syntézy odpovědí) a metody, které jsou specifické pro jediný typ otázky.

5.2.1 Rozklad otázky a syntéza odpovědí

Mnoho otázek typu „PUZZLE“, včetně „PŘED A PO“ a většina otázek typu „RÝMOVAČKY“ („*Rhyme Time*“) vyžadují pro úspěšné vyřešení rozklad. Na příklad odpověď k otázce:

PŘED A PO: Hvězda filmu „Jerry Maguire“, která zároveň automaticky udržuje rychlost ve vašem vozidle. („Tom Cruise control“ - česká verze: „Tom tempomat“) (6)

je vymyšlená fráze, kterou s největší pravděpodobností nenajdeme v žádném referenčním korpusu. Odpověď je v tomto případě tvořena kombinací výsledků, které jsou získané pomocí dvou oddělených hledání. Obecně funguje rozklad otázky a syntéza odpovědí na podobném principu (viz Illustration 1). Obecně určuje stylizovaná forma otázky způsob, jakým je potřebné provézt rozklad.

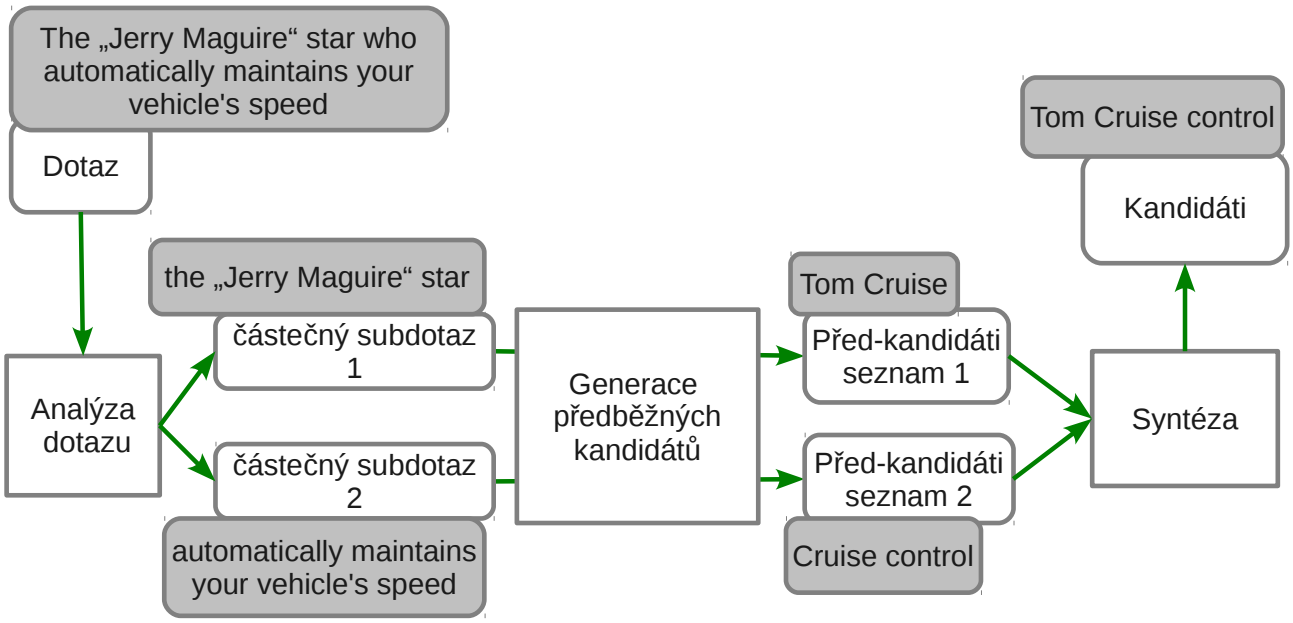

Illustration 1: Rozklad toku. Obecný tok pro proces rozkladu na příkladu "Tom Cruise control". Přibližně 100 před-kandidátů se vygeneruje pro každý částečný subdotaz. Sestaveno na základě ilustrace z [2] (Pro přesnou názornost ponechávám znění otázky v originále)

Konkrétně rozkladu otázek typu „PUZZLE“ se provádí na základě shody manuálně vyvinutých Prologových pravidel a struktur predikát-argument (PAS) k identifikaci subdotazů. K rozpoznaným částečným subdotazům (*SubQuestionSpans*) se následně vygenerují kandidáti na správnou odpověď, označených jako předběžní kandidáti nebo před-kandidáti, načež se tito před-kandidáti spojí v další fázi (syntéza odpovědí) a to například překryvem u typu „PŘED A PO“ nebo zvukovou shodou (rýmem) u otázek typu „RÝMOVAČKY“.

5.2.2 Omezení a slovní hříčky (*constraints and puns*)

Společnou vlastností otázek v soutěži Jeopardy! je, že obsahují nápovědi k správně odpovědi ve formě lexikálních omezení (*lexical constraints*) a slovních hříček (*puns*). Lexikální omezení jsou nejčastěji obsažené v názvu kategorie (ale někdy také v textu samotné otázky). Obvykle se jedná o požadavky na délku slova nebo výrazu odpovědi, nebo omezení typu „začíná na“, „končí na“, „obsahuje... (určitá písmena nebo posloupnost písmen)“, nebo vytváření rýmu nebo další lexikální vlastnosti. Odhadem 13% otázek v soutěži Jeopardy!, včetně asi 35% otázek typu „DEFINICE“, obsahují jednu nebo více takovýchto omezení. Slovní hříčky jsou mnohem méně zastoupené – asi v 1% případů – nicméně jsou mnohem víc náročnější na zpracování a snad i zajímavější právě z důvodu omezeného vývoje v oblasti počítačového humoru. Takovéto vlastnosti se nepoužívají při generování kandidátů na odpověď, ale pro ohodnocení existujících kandidátů na odpověď na základě stupně shody s rozpoznanými elementy slovních hříček. V principu jsou tyto nápovědi použity pro jakýkoliv typ dotazu, ale prakticky se nejčastěji uplatňují v Standardních Jeopardy! Otázkách (*Standard Jeopardy! Questions*) a to jak ve formě doporučení v procesu konstrukce odpovědi, tak jako pojistka jedinečnosti v případě nejednoznačné odpovědi.

5.2.3 Lexikální omezení

Lexikální omezení může být instancí typu:

ZAČÁTKY A KONCE NA „A“: Tato země sehrála roli prostředníka v roce 1981 pro zajištění propuštění 52 Američanů držených v Íránském zajetí. („Alžírsko“) (7)

nebo může omezovat odpověď částečně:

POUZE JEDNA SAMOHLÁSKA – pro zachování smyslu, ponechávám v originále: Proverbially, you can be „flying“ this or be this „and dry“. („high“) (8)

Dotaz v příkladu (8) je zajímavý z několika pohledů – obsahuje lexikální omezení, není rozložitelný a je to otázka typu „FITB“ („doplňovačka“, *Fill-in-the-blank*).
Lexikální omezení mohou být také vágní a nejednoznačná, například citované výrazy v kategoriích naznačují, že tento citovaný text bude částí každé odpovědi, ale není vždy jasné (ani pro lidi), zda tento text je na začátku odpovědi, nebo jestli je obsažen na začátku každého slova odpovědi, nebo jestli se bude vyskytovat na začátku jakéhokoliv slova v odpovědi.

Syntax kategorií, které nesou lexikální omezení, nemusí být vždy v angličtině (například: THE L“ONE“LIEST NUMBER ~ Nej“osa“mělejší číslo, GOING INTO O__T) nebo je problematická s ohledem na hranice slov (JEDNO O, PAK 2 O – v originále: ONE O, THEN 2 O's). Při testování původního algoritmu na nových otázkách, s kterými se předtím Watson nesetkal se ukázalo, že existující rozpoznávač (*recognizer*) nedokáže identifikovat některá omezení vyjádřena nečekaným způsobem. Nicméně díky tomu, že omezení se obvykle nachází v názvu kategorií, společné vzorce z odpovězených dotazů v kategoriích jsou užitečné pro odvození omezení v neúspěšných případech.

5.2.4 Omezující objekty (*constraint objects*)

Omezující podmínky jsou zpracovávané samostatnou komponentou (*Contrainer component*), která pozůstává z tříd omezení (*constraint classes*) a řídicí infrastruktury (*management infrastructure*). Každá třída omezení zpracovává jiný typ omezujících podmínek a zvládá následující funkce:

- rozpoznávání a vytváření instancí konkrétní třídy omezení z dotazu

- odvození třídy omezení ze zodpovězených otázek v jedné a té samé kategorii
- testování kandidátů na správnou odpověď oproti splněným omezujícím podmínkám

Třídy omezení jsou sestavené manuálně tak, aby pokryly co největší množství často vyskytujících se lexikálních omezení. Například jedna třída pozůstává z omezení, která specifikují délku odpovědi co do počtu znaků, zatímco jiná třída obsahuje omezení, která blíže specifikují požadavky na některou z podčástí odpovědi (subdotazy). Instance vytvořené z objektů omezení se mohou spojovat. Například z kategorie „SLOVA ZAČÍNAJÍCÍ NA „W“ NA 10 PÍSMEN“ se vytvoří dva omezující objekty. Jeden propustí pouze odpovědi složené z deseti písmen a druhý propustí pouze odpovědi začínající na „w“. V konečném testování možných správných odpovědí (před-kandidátů) musí být splněny současně obě omezující podmínky. Watson pracuje s několika desítkami různých tříd omezení – Table 4:Příklady lexikálních omezení uvádí některé z těchto tříd spolu s kategoriemi, které je vyvolá a odpovědmi, které touto podmínkou projdou.

Table 4: Příklady lexikálních omezení

Třída omezení	Příklad kategorie (AJ)	Příklad odpovědi
Aliterace (*Alliteration*)	ALLITERATIVE ARTISTS	Pablo Picasso
ChybějícíVýraz (*Blank*)	ON____	On **Deck**
ZdvojenéPísmeno (*DoubleLetter*)	DOUBLE „H“	Fis**hh**ook
SpolečnýZávěr (*CommonEnding*)	-OID	Cellul**oid**
OpakováníPísmen (*NLetterRepetitions*)	„U“2	F**utu**res
Rýmování (*Rhyming*)	RHYMES WITH JOCK	Frock
RozděleníŘetezců (*SubStringDisjunkction*)	„CHURCH“ AND „STATE“	Churchkey
PočetSlabik (*SyllableCount*)	MONOSYLLABLES	Ain't

Všechny příklady uvedené v tabulce jsou odvozené z kategorie, ale v některé instance jsou někdy obsažené i v samotném textu otázky, např.: „Stejně jako mláďata čeledi psovitých i mladé potkany se nazývají tímto *jednoslabičným* pojmem.“
Proces zpracování omezujících podmínek (*Constrainer process*) je spuštěn až po analýze dotazů a to tak, že dotaz je srovnán s každým omezujícím objektem kvůli tomu, zda má být aktivovaný nebo nikoliv. Počet aktivních omezujících předmětů se uloží do tzv. souboru aktivních omezení (*ActiveConstraintSet*). Každý před-kandidát je Watsonem testován vůči tomuto souboru.

5.2.5 Slovní hříčky

Slovní hříčky v otázkách soutěže Jeopardy! se nevyskytují zrovna často a pokud ano, je to čistě z důvodu pobavení. Většinou jde o to podat obyčejné výrazy v chytlavé formě. Obvykle není potřeba jim přímo rozumět nebo je zpracovat jinou formou (než jen „pohrát se“ s písmeny ve slově). Na druhou stranu, slovní hříčky v otázce jsou obvykle jasnou nápovědou k odpovědi. Nejvíce se opakující slovní hříčky, které byly v IBM rozpoznány – a jediné, které se pokusily řešit – jsou hříčky ve formě citovaných frází, které modifikují jádro dotazu. V IBM identifikovali čtyři druhy

podtříd slovních hříček:

- PODŘĚTEZEC (*Substring*): SLOGANY: Se sloganem „Chytni vlnu“ („*Catch the wave*“) byl v roce 1987 uvedený Maxem Headroomem tento „nový“ („*new*“) nealkoholický nápoj. („Nová Kola“, v originále „New Coke“ - slovo „new“ tedy neoznačuje jen vlastnost uvedeného výrobku – tedy že se jedná o novinku – ale zároveň je součástí odpovědi)
- SYNONYMA: GEOLOGIE – klíčová slova budu uvádět v angličtině, kvůli zachování významu: Kvůli svému lesku dali němečtí horníci tomuto „pleasant“ (příjemnému) kameni své jméno, které znamená „spark“ (jiskra). (odpovědí je slovo „gneiss“ s výslovností [najs], kterého homofonum zní jako anglické slovo „nice“, což je synonymem slova „pleasant“, tedy příjemný)
- INSTANCE: *F*: Opera z roku 1900 dala křídla této populární „hmyzí“ skladbě od Rimsky-Korsakova. („Flight of the Bumblebee“ - překlad: Let čmeláka)
- ASOCIACE: ZNÁM JACKA: Jednalo se o TV debut Marilyn Monroe v roce 1953 v této „lakomé“[15] komediální TV show. („Jack Benny“)

Výraz slovní hříčky a před-kandidáta (správné odpovědi) jsou lemmatizovány a poté vzájemně porovnávány vůči shodě – pokud se jedná o víceslovný výraz, srovnávají se jednotlivá slova stejně tak jako celý řetězec slov (celý víceslovný výraz). Operace srovnávání také obsahuje homofona. Pro typ „PODŘETĚZEC“ je výraz slovní hříčky podřetězcem odpovědi. Pro „SYNONYMA“ je výraz slovní hříčky synonymem odpovědi – tady se využívá WordNetu. U typů „INSTANCE“ je odpověď instanční nebo subtypem slovní hříčky – s použitím YAGO TyCor. U typu „ASOCIACE“ existuje silná asociace (která ale neodpovídá žádnému z předešlých případů) mezi pojmem odpovědí a pojmem slovní hříčky. Testování této asociace bylo měřeno stupněm korelace mezi dvěma pojmy s použitím korpusu n-gramů extrahovaných z kolekce anglických textů obsahujících články z encyklopedií a informačních agentur. Tento stupeň asociace se pak srovnává s empiricky stanovenou hladinou, která rozhoduje o významnosti této asociace se slovní hříčkou.

Bližší informace ohledně specifických dotazů a metod jejich zpracování jsou uvedeny v článku „Special Questions and techniques“ [2].

5.3 QSekce (QSections)

QSekce je označení určité souvislé části textu otázky (výjimečně kategorií), která má důležitou funkci v interpretaci dotazu. Podobně jako v případě QTříd jsou QSekce identifikovány pomocí pravidel Prologu nad strukturami typu predikát-argument (PAS) nebo pomocí regulárních výrazů nad textem. V některých případech jsou QTřídy a QSekce identifikovány současně, jelikož jejich existence a funkce jsou vzájemně závislé. Mezi nejdůležitější QSekce patří:

- Lexikální omezení – fráze jako „toto slovo na 4 písmena“, které by nemělo být použito v dotazu, ale které je kritické pro výběr správné odpovědě.
- Zkratky – výraz v otázce, který je interpretován jako zkratka, která je asociovaná s možným

15 Šlo o známou americkou show, která se vysílala několik desítek let. Show běžela nejdříve pouze v rozhlasu, později se přesunula do televize. Jednalo se o situační komedii s hlavní postavou skrblíkem Jackem Bennym – proto slovo „lakomý“ tvořilo nápovědu k názvu pořadu. Vtip vystihující povahu Jacka Benny z jedné epizody:
Jack Benny se vracel domů, když byl přepaden – zloděj jej nejdříve poprosí o zapalovač, aby si mohl připálit cigaretu a poté na něj zloděj vybalí: „Ani hnout, toto je přepadení! A teď naval peníze nebo život!“ Benny strnule stojí bez reakce. Již v této fázi zazní v obecenstvu (které zná povahu hlavní postavy) smích. Zloděj zopakuje: „No tak, hni s sebou! Povídám peníze nebo život!“ Načež mu Benny odsekne: „Přemýšlím nad tím!“
Více na: https://en.wikipedia.org/wiki/Jack_Benny#.22Your_money_or_your_life.22

hledaným významem. Pro otázky z QTřídy „ZKRATKY“ je jedna z QSekcí „ZKRATKY“ dále identifikována jako zkratka, které význam je hledanou odpovědí.

- Částečný subdotaz (*SubQuestionSpan*) – v případě, kdy může být otázka rozložena na dvě nebo více disjunktních sekcí, z kterých každá o sobě odkazuje na nebo naznačuje otázku, jsou tyto sekce označené jako částečné subdotazy (*SubQeustionSpan*). Týká se to některých složených Faktoidů (str. 4, poznámka pod čarou č. 3) jako „PŘED A PO“, „RÝMOVAČKY“ a dvojitých definic jako je první příklad typu „DEFINICE“ v Table 3:QTřídy (QClasses).
- McAnswer (nejsem si jistá překladem) – Řetězce (obvykle tří), které reprezentují možnosti odpovědi v otázce typu „VÍCE MOŽNOSTÍ“ jsou označené jako McAnswer QSekce[16].
- „FITB“ - „doplňovačky“ (*Fil-in-the-blank*) – označení pro řetězec, který je přidružený k jádru (tj. text, který vytváří termín nebo výraz, který je doplněný jádrem).

5.4 Hodnocení

Vyhodnocení (testování) přesnosti proběhlo vůči manuálně anotovanému zlatému standardu, který pozůstával z přibližně 3500 otázek. Celková hodnota F u všech QTříd (str. 12, vzorec (4) dosáhla 0,637[17] s tím, že mnoho QTříd má tuto hodnotu vyšší (viz Table 5:Vyhodnocení klasifikace dotazů (Question Classification)).

Table 5: Vyhodnocení klasifikace dotazů (Question Classification)

Qtřída (*QClass*)	**Přesnost (*Precision*)**	**Pokrytí (*Recall*)**	**F_1**
DEFINICE	0,508	0,487	0,497
KATEGORIE-RELACE	0,644	0,806	0,716
FITB	0,676	0,711	0,693
ZKRATKY	0,728	0,806	0,765
PUZZLE	0,977	0,525	0,683
ETYMOLOGIE	0,886	0,600	0,716
SLOVESA	0,796	0,811	0,804
PŘEKLAD	0,885	0,590	0,708
ČÍSLA	0,842	0,471	0,604
VZÁJEMNÉ VZTAHY	1,000	0,652	0,789
VÍCE MOŽNOSTÍ	0,650	0,684	0,667
DATA	0,692	0,818	0,750

16 Dle wiki, je Mc obdobou Mac mající význam *son of* (syn ...) – je tedy možné, že označení McAnswer má naznačit, že odpověď je jedna z množiny možných odpovědí („dceřinka“); https://cs.wikipedia.org/wiki/MC

17 Z článku není jasno, jak bylo F_{all} vypočítáno – o aritmetický průměr se nejedná.

Qtřída (*QClass*)	**Přesnost (*Precision*)**	**Pokrytí (*Recall*)**	**F_1**
Celkem	0,646	0,629	0,637

QTřída s nejnižší hodnotou detekce F je „DEFINICE“ možná právě proto, že je to jedna ze subjektivních klasifikací. Chyby v detekci „DEFINICE“ často nevadí, protože otázky tohoto typu jsou podobné otázkám typu FAKTOID a tím pádem mohou být otázky typu „DEFINICE“ zodpovězeny správně i při nesprávném zařazení do QTřídy. Také QTřída „PUZZLE“ má nízkou hodnotu pokrytí (*recall*), co je způsobeno zejména tím, že v otázkách se vyskytují typy puzzle (hádanek apod.) které nejsou moc časté a nebylo proto efektivní vyvíjet pro ně speciální detekční mechanizmy. Při hodnocení celého systému analýzy dotazů byl porovnáván celý systém Watson s konfigurací, u které byly úplně odstraněny komponenty QTřídy a QSekce = tedy jenom systém se základními pravidly (*baseline system*). Toto srovnání je uvedeno v Table 2:Vyhodnocování účinnosti detekce LATu. Srovnání třetího a čtvrtého řádku ukazuje, že se zapojením systému pro detekci QTříd a QSekcí se zvýšila správnost odpovědí Watsona o dalších 2,9%. Tento nárůst je statisticky významný podle McNemarova testu na hladině $p < 0,01$ (a tedy zvýšení úspěšnosti není náhodné, a je způsobeno systémy pro detekci QTříd a QSekcí).

V systémech, které jsou vyvinuté obecně pro zodpovídání dotazů, se často otázky reprezentují jako grafy buďto sémantických relací v analyzovaném úseku (*parse*), nebo PAS, nebo jako hluboké (vnořené) sémantické relace (*deep semantic relations*) v ručně vytvořených ontologiích. Watson využívá obou přístupů.

Koncept jádra dotazu nemá úplně přesnou definici a možná nejsrozumitelnější rozpravu na téma detekce jádra dotazu lze najít v článku Bunescu a Huang (odkaz na zdroj lze najít v originále článku „Question Analysis: How Watson reads a clue“ [1]). Bunescu a Huang definují jádro dotazu jako „množinu všech maximálních jmenný frází v otázce, které se spoluodkazují na odpověď (koreferují)“. Dále uvádí, že slova jádra dotazu mohou ovlivnit výběr typu odpovědi. Ve své práci neuvažují samostatný koncept LATů a tím pádem neuvažují případy, kdy slovo není jádrem dotazu a přitom může stále ovlivnit výběr typu odpovědi. Pro Watsona je toto rozlišení mezi jádrem dotazu a LATem nezbytné kvůli jeho heterogenní množině algoritmů pro hodnocení odpovědi. Některé jsou založené na souladu odpovědi s jádrem dotazu, jiné využívají LAT pro ověření správné volby typu odpovědi. Například v otázce „Byl zvolen za prezidenta v roce 1868.“ je slovo „prezident“ evidentně LATem, ale určitě nechceme, aby jej Watson považoval za jádro odpovědi, jelikož by bylo chybné sladit jej s odpovědí (neočekáváme, že by se odpověď k této otázce objevila jako rozvíjející větný člen k slovesu „volit“).

Mnoho systémů pro odpovídání na otázky používají pro analýzu dotazů identifikaci sémantického typu odpovědi na základě pevných ontologií známého typu. Tento přístup není v případě Jeopardy! soutěže příliš praktický. Místo toho se u Watsona používá oddělených přístupů, kdy se v rámci analýzy dotazu nejdříve identifikuje LAT, ale komponenty vynucených typů (*type coercion components*) určují sémantiku. Podobným přístupem se zabývá práce Pinchaka a Lina (odkaz na zdroj lze najít v originále článku „Question Analysis: How Watson reads a clue“ [1]), kteří používají přímo slova z otázek pro určení typu odpovědi, i když mají jeden specifický algoritmus pro evaluaci před-kandidátů na správnou odpověď postaveného na bázi četnosti výskytu ve stejném kontextu v korpusu, zatímco Watson identifikuje LATy nezávisle na tom, jak jsou použité a až potom použije sadu algoritmů pro vynucené typy (*type coercion algorithms*).

6 Zdroje

[1] : Lally A. et al. *Question analysis: How Watson reads a clue.* IBM J. RES. & DEV. NO. 3/4. Paper 2. VOL. 56. May/July 2012. URL <http://researcher.watson.ibm.com/researcher/files/us-heq/W%284%29%20QUESTION%20ANALYSIS%2006177727.pdf>. Cit. 12.5.2013.

[2] : Prager J.M. et al. *Special Questions and techniques*. IBM J. RES. & DEV. NO. 3/4. Paper 11. VOL. 56. May/July 2012. URL <http://researcher.watson.ibm.com/researcher/files/us-heq/W%2813%29%20SPECIAL%20QUESTIONS%2006177732.pdf>. Cit. 1.6.2013.

[3] : Leccos – Gramém. URL<http://leccos.com/index.php/clanky/gramem>. Cit. 23.05.2013.

[4] : Erhart, A. *Úvod do jazykovědy*. Brno: Masarykova Univerzita, 2001. 1. vyd. ISBN 8021026693

[5] : McCord M.C. *Slot Grammars*. Association for Computational Linguistics. American Journal of Computational Linguistics. Number 1. Volume 6. January - March 1980. URL <http://acl.ldc.upenn.edu/J/J80/J80-1003.pdf>. Cit. 23.05.2013.

[6] : McCord M.C. *Using Slot Grammar*. New York: IBM Research Division. IBM Research Report. March 24, 2010. URL <http://domino.research.ibm.com/library/cyberdig.nsf/papers/FB5445D25B7E3932852576F10047E1C2/$File/rc23978revised.pdf>. Cit. 23.05.2013.

[7] : Noun phrase. Wikipedia. 27 May 2013. URL <https://en.wikipedia.org/wiki/Noun_phrase>. Cit. 25.5.2013.

[8] : Wh-pronoun. An Encyclopedia of Word Grammar and English Grammar. 2002. URL <http://www.phon.ucl.ac.uk/home/dick/enc2010/frames/frameset.htm>. Cit. 23.05.2013.

[9] : Feature structure. Wikipedia. 11 March 2013. URL <https://en.wikipedia.org/wiki/Feature_structure>. Cit. 25.5.2013.

[10 : Precision and recall. Wikipedia. 15 May 2013. URL <https://en.wikipedia.org/wiki/Precision_and_recall>. Cit. 1.6.2013.

[11] : Sensitivity and specificity. Wikipedia. 31 May 2013. URL <https://en.wikipedia.org/wiki/Sensitivity_and_specificity>. Cit. 1.6.2013.

Další zdroje o systému Watson:
Deep Parsing in Watson: http://researcher.watson.ibm.com/researcher/files/us-heq/W(5)%20DEEP%20PARSING%2006177729.pdf
Automatic knowledge extraction from documents: http://www.andrew.cmu.edu/user/ooo/watson/05%20automatic%20knowledge%20extration.pdf
Typing candidate answers using type coercion: http://researcher.watson.ibm.com/researcher/files/us-heq/W(9)%20TYPE%20COERCION%2006177730.pdf
Relation extraction and scoring in DeepQA: http://researcher.watson.ibm.com/researcher/files/us-heq/W(11)%20RELATION%20EXTRACTION%2006177734.pdf
Structured data and inference in DeepQA: http://researcher.watson.ibm.com/researcher/files/us-heq/W(12)%20STRUCTURED%20DATA%2006177725.pdf
Identifying implicit relationships: http://researcher.watson.ibm.com/researcher/files/us-heq/W(14)%20IMPLICIT%20RELATIONSHIPS%2006177721.pdf
Fact-based question decomposition in DeepQA: http://researcher.watson.ibm.com/researcher/files/us-heq/W(15)%20QUESTION%20DECOMPOSITION%2006177726.pdf
A framework for merging and and ranking of answers in DeepQA: http://www.andrew.cmu.edu/user/ooo/watson/14%20a%20framework%20for%20merging%20and%20ranking%20answers.pdf
Making Watson fast: http://www.andrew.cmu.edu/user/ooo/watson/15%20making%20watson%20fast.pdf