PlayCoref

podrobná specifikace
(Lenka Studničná, verze 24.9.2009)

1. Popis implementované hry

Základní algoritmus hry pro hráče:

1. Registrace a přihlášení na herní portál, výběr hry, výběr levelu (velikost dokumentu) a volba, zda čekat na protihráče, nebo zda hrát proti automatickému soupeři.
2. Potvrzení zahájení hry.
3. Hráči se objeví cca 3 věty ze vstupního dokumentu. Hráč má možnost kdykoliv k již zobrazeným větám odhalit větu další (kliknutím na příslušné tlačítko), dokud není zobrazen celý dokument. Všechny odhalené věty jsou viditelné zároveň a je umožněna práce se všemi z nich.
4. V odkrytých větách jsou zvýrazněna „aktivní“ slova, která může hráč spojovat do koreferenčních párů. Kterýkoliv vytvořený pár může opět smazat (kliknutím na tlačítko *odstranit* a následným kliknutím na vytvořenou hranu mezi dvěma slovy). Celkový počet vytvořených párů je zobrazen. Zobrazován je průběžně i počet párů vytvořených soupeřem (případně ještě počet jeho odkrytých vět), jiné informace o protihráči či možnost jakékoliv komunikace s ním hráč nemá.
5. Po skončení hry je zobrazeno výsledné skóre obou hráčů, vyhlášen vítěz, zobrazeny další informace (třeba počty v kolika párech se hráči shodli navzájem a v kolika ne).

Průběh registrace hráče, výběru hry atd. je společný se Shannonovou hrou (tedy v tomto respektuji její specifikaci)

Základní algoritmus pro zpracování dat:

Před hrou

1. Cílem je získat co nejvíce paralelních dat pro každý dokument. Ty jsou proto označeny ID čísly a pro každou hru je zvolen dokument s nejnižším ID takový, že s ním žádný ze zúčastněných hráčů ještě nehrál. Alternativně je možné přiřazovat dokumenty podle jejich obtížnosti a spolehlivosti hrajících hráčů.
2. Vstupní data (PML dokument) jsou předzpracována: podle tagu jednotlivých slov jsou vybrána slova, která mohou být součástí koreferenčních párů. Tato se pak hráčům jeví jako aktivní pro tvoření párů.

Po hře

3. Porovnání párů vytvořených hráči - mezi hráči a ve vztahu s referenčním/automaticky vytvořeným párováním. Spočítání skóre pro hráče. Podle shody může být přiřazeno i určité skóre spolehlivosti dat pro každý získaný pár.
4. „Zorientování“ hran vztahů v označených párech podle pořadí slov v dokumentu a případně jejich spojení do koreferenčních řetězců. Uložení dat. Z uložených dat je možné vyexportovat výstupní soubory.

Grafické prostředí hry

- Zobrazení vět s „aktivními“ slovy, která lze spojovat do párů (kliknutím na jedno slovo a následně na druhé, slova jsou si v páru rovnocenná). Vybraný pár je pak dále barevně odlišen nebo spojen čárou.
- Zobrazení odpočtu času, počet už zobrazených (anebo ještě nezobrazených) vět a počet označených párů hráče i jeho soupeře

- Tlačítka
  - pro zobrazení další věty
  - pro smazání vytvořené hrany
  - pro ukončení hry (vzdání se)

Bodové ohodnocení hry

Výpočet skóre je přesně popsán v článcích, stručně:
$player\_A\_score = w_1 \cdot ICA(A, automatic\_coreference\_resolution) + w_2 \cdot ICA(A, B) + w_3 \cdot N(A)$

Hra pro jednoho hráče

Kvůli (alespoň zpočátku) nízkému počtu reálných protihráčů je nutné umožnit hru pro jednoho hráče. Problém je ve způsobu ohodnocení tohoto hráče. Návrhy na řešení:

1. Hrát pouze proti „zlatému standardu“, tj. pouze porovnání s daty už spolehlivě oanotovanými koreferencemi (PDT). Pak ale hra neposkytne žádná nová data.
2. Počítat skóre jen na základě shody hráče s automaticky vygenerovanými páry. Bohužel je náchylné k podvádění a naopak bude znevýhodňovat hráče, jehož párování je lepší než automatické.
3. Použít druhý odlišný algoritmus automatického označování párů a zacházet s ním jako s druhým reálným hráčem.

Sledovaná a ukládaná data o partiích

- jazyk hry
- hráči
- odkaz na dokument, se kterým se hrálo
- počet vět v dokumentu, počet odkrytých vět jednotlivými hráči, časový limit partie
- vytvořené páry

Sledovaná a ukládaná data o označených párech

- čas
- první označené slovo
- druhé označené slovo (přestože pro skóre pořadí nerozhoduje)
- příznak, zda byl později pár hráčem odstraněn

2. Požadavky na funkčnost

Předpokládám využití a podřízení se už implementovaným částem herního portálu, jak jsou popsány ve specifikaci k Shannonově hře.

(Správa uživatelů a nainstalovaných her, řízení průběhu hry, architektura (vrstvy))

Import vstupních dat

Dokumenty ve formátu PML. Relevantními informacemi pro hru je rozčlenění dokumentu na věty (s), slova (form) a jejich tagy (tag).
Velikost dokumentu by měla být omezena, aby zkušený hráč stihl zobrazit během hry nejlépe celý dokument a tedy byl nucen hledat i „méně nápadné“ koreference. Cca 15 vět pro limit 5 minut?

Export sesbíraných dat

Hlavními daty je databáze dvojic slov, která hráč označil jako koreferenci. Ve dvojicích můžeme podle pořadí slov ve větě označit antecedent-anafor. Několik dvojic může být seskupeno do jednoho koreferenčního řetězce.
Forma uložení výstupních dat, varianty:

- Obohacení vstupních souborů her o odkazy na antecedent u každé anafory a přidání informací o hře (hrách), při které byla data získána.
- Vlastní XML soubor s pouze vypsanými koreferenčními řetězci (formou odkazů na původní dokument) a informacemi o hře a hráči. Předpokládám, že tento přístup usnadní následné zpracování získaných paralelních dat pro jeden dokument.

3. Návrh databáze

- **Games** (hry dostupné na herním portálu; společné pro celý portál, převzato z implementace Shannonovy hry)

název	**datový typ**	**popis**	**key**
id	SERIAL	identifikátor hry	PK
name	VARCHAR(32)	název hry	
description	TEXT	popis hry	
active	BOOLEAN	zda je hra zobrazována	
url	VARCHAR(255)	umístění hry	
image	BYTEA	data obrázku	

- **Players** (zaregistrovaní hráči; společné pro celý portál, převzato z implementace Shannonovy hry)

název	**datový typ**	**popis**	**key**
id	SERIAL	identifikátor uživatele	PK
email	VARCHAR(64)	e-mailová adresa uživatele	
login	VARCHAR(16)	unikátní jméno uživatele	
password	VARCHAR(32)	heslo (MD5)	
gender	CHAR(1)	pohlaví uživatele	
birth_year	INTEGER	rok narození uživatele	
language	CHAR(2)	mateřský jazyk uživatele	

- **Pool** (hráči čekající na zápas; společné pro celý portál, převzato z implementace Shannonovy hry)

název	**datový typ**	**popis**	**key**	
game_id	INTEGER	odkaz na hru	FK (Games)	PK
player_id	INTEGER	odkaz na hráče	FK	

			(Players)	
level	INTEGER	uživatelem zvolená obtížnost/skupina dle délky dokumentu		
language	CHAR(2)	uživatelem zvolený jazyk hry		
last_update	TIMESTAMP	čas posledního přístupu do poolu		

- **Matches** (jednotlivé zápasy PlayCoref)

název	**datový typ**	**popis**	**key**
id	SERIAL	identifikátor partie	PK
game_id	INTEGER	odkaz na hru	FK (Games)
document_id	INTEGER	odkaz na dokument	FK (Players)
created	TIMESTAMP	čas vytvoření zápasu	
started	TIMESTAMP	čas startu zápasu	
finished	TIMESTAMP	čas ukončení zápasu	
players	INTEGER	počet hráčů	
level	INTEGER	obtížnost hry	

- **Matches_Players** (informace o výsledcích hráčů)

název	**datový typ**	**popis**	**key**	
match_id	INTEGER	odkaz na zápas	FK (Matches)	PK
player_id	INTEGER	odkaz na hráče	FK (Players)	
status	VARCHAR(8)	status zápasu		
score	INTEGER	dosažené skóre hráče		
win	BOOLAEN	výhra hráče		

- **Documents** (vstupní dokumenty)

název	**datový typ**	**popis**	**key**
id	SERIAL	identifikátor dokumentu	PK
sentences	INTEGER	počet vět v dokumentu	
language	CHAR(2)	jazyk dokumentu	

- **Sentences** (vstupní věty)

název	**datový typ**	**popis**	**key**

document_id	INTEGER	odkaz na dokument	FK (Documents)	PK
position	INTEGER	pořadí věty v dokumentu		
words	INTEGER	počet slov ve větě		

- **Words** (vstupní slova)

název	**datový typ**	**popis**	**key**	
document_id	INTEGER	odkaz na větu	FK (Sentences)	PK
sentence_position	INTEGER			
position	INTEGER	pořadí slova ve větě		
token	VARCHAR(32)	slovo		
tag	VARCHAR(16)	morfologická značka		
active	BOOLEAN	slovo je pro hráče aktivní		

- **Pairs** (hráčem označené koreferenční páry)

název	**datový typ**	**popis**	**key**	
match_id	INTEGER	odkaz na zápas		PK
player_id	INTEGER	odkaz na hráče		
1_document_id	INTEGER	odkaz na první hráčem vybrané slovo	FK (Words)	
1_sentence_position	INTEGER			
1_word_position	INTEGER			
2_document_id	INTEGER	odkaz na druhé hráčem vybrané slovo	FK (Words)	
2_sentence_position	INTEGER			
2_word_position	INTEGER			
created	TIMESTAMP	čas vytvoření dvojce		
deleted	BOOLEAN	zda hráč smazal		

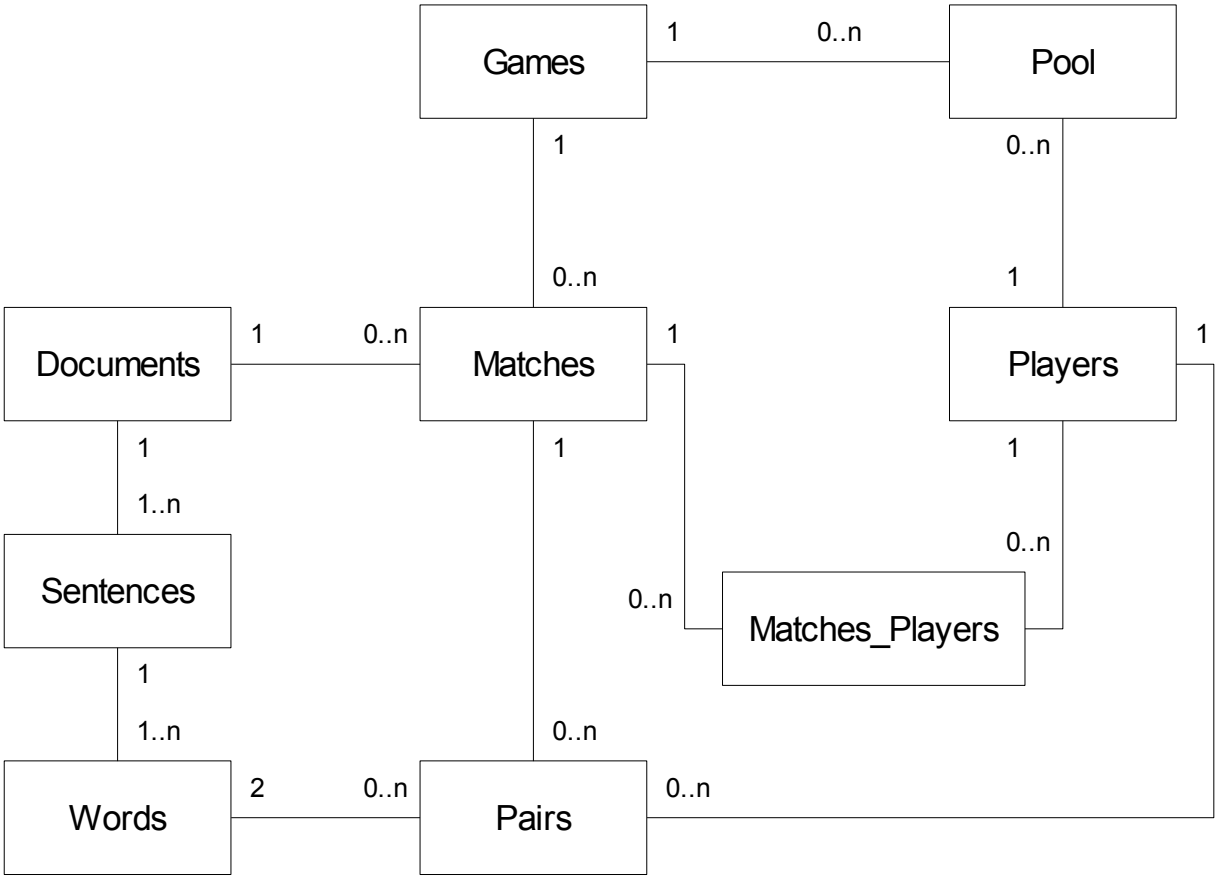
Tabulky jsou navrženy tak, aby byla minimalizována náročnost dotazů na databázi během zápasu. Dotazy mimo hru mohou být na zpracování náročnější (rozsáhlejší skládání tabulek), konkrétně se jedná zejména o dotazy nutné pro export užitečných dat z databáze (např. všechny dosud označené koreference v jednom dokumentu).

Alternativy:

- Pokud budeme schopni přiřadit vstupním dokumentům jejich obtížnost, je možné přidat do **Documents** atribut **level**.
- Pokud se některé často používané přístupy do databáze ukážou jako příliš náročné, mohou být vytvořeny přídavné pomocné tabulky s předzpracovanými daty.
  - příklad (pouze ukázkový, ne potřebný) – zvláštní tabulka s počty odehraných her pro

jednotlivé hráče. Počty jsou aktualizovány po každé hře a pouze méně často (cron) kontrolována konzistence mezi počtem her zde a počtem získaným z tabulky Matches_Players

Vazební vztahy:



4. Komunikace mezi Flashovou a PHP vrstvou

Zprávy zasílané mezi vrstvami mohou být prakticky shodné se Shannonovou hrou (viz její specifikace), změny jsou nezbytné pouze v případech, kdy jsou přenášena data o označených dvojicích a výsledcích – je nutné zvolit jiný formát. Konkrétně se jedná o:

- návratové hodnoty **words** u **game-start**
  - Formát může být zachován, nebudou ale samozřejmě vynechávána žádná slova.
- návratové hodnoty **results** u **game-status**
  - Potřebujeme u hráčů udržovat jimi označené dvojice
- v **game-guess** místo **position** a **word** bude **position1** a **position2**
- v **game-results** nebude použita položka **sentence** a formát **results** bude změněn, jako to je u **game-status**

A přidat vhodné změny pro umožnění hry jednoho hráče.

5. Idey na rozšíření

Dlouhodobé skóre – bude zahrnuto

Za podstatné pro zajištění kvality výstupních dat považuji uchovávání určitého dlouhodobého skóre pro každého hráče (pro PlayCoref a zvlášť i pro ostatní hry), které bude záviset na počtu odehraných her (zkušenost hráče) a dosažených výsledků v nich (spolehlivost hráče). U hráčů s nízkým skóre nemusí být jejich data dále vůbec zpracovávána, nebo alespoň může být nějak snížena jejich relevance – to bude užitečné při následném zpracování paralelně získaných dat.

Víceslovné výrazy – pravděpodobně bude zahrnuto

Umožnění práce s víceslovnými výrazy (skupina slov je pro hráče prezentována jako jediný aktivní celek). Toto by si ale vynutilo změnu formátu vstupních dat.

Osobní výsledky

Na portál přidat stránku, kde budou statistiky přihlášeného hráče – kolikrát kterou hru hrál, kolik má vítězství, kolik času strávil hraním apod.