

Rudolf Rosa
rosa@ufal.mff.cuni.cz

Chatbot Pohádkové dítě

Fairytales Child Chatbot

Univerzita Karlova v Praze
Matematicko-fyzikální fakulta
Ústav formální a aplikované lingvistiky



ITAT 2014, Demänovská dolina, 26. září 2014

Pohádkové dítě

- Chatbot v konzoli (standardní vstup a výstup)
 - ve stylu chatbota ELIZA (Joseph Weizenbaum)
 - ale pro češtinu
- Chatbot představuje pohádkychtivé dítě
 - chce po uživateli, aby mu vyprávěl pohádku
 - analyzuje vstup od uživatele
 - snaží se vhodně reagovat (otázkou)
- Na konci prezentace bude demonstrace
 - přemýšlejte nad svou oblíbenou pohádkou!

1. Morfologická analýza vstupu

- *Uživatel:* Bylo jednou dvanáct mladých princezen.
- tokenizace, lematizace, tagging

token	lemma	tag
▪ Bylo	být	VpNS - - - XR - AA - - -
▪ jednou	jednou-2	Db - - - - - - - - - - -
▪ dvanáct	dvanáct`12	Cn - S1 - - - - - - - - -
▪ mladých	mladý	AAFP2 - - - - 1A - - - -
▪ princezen	princezna	NNFP2 - - - - - A - - - -
▪ .	.	Z : - - - - - - - - - - -

2. Výběr šablony odpovědi

- *Uživatel:* Bylo jednou dvanáct mladých princezen.
 - **Co je to** <podstatné jméno>?
 - Co jsou to princezny?
 - **Jaké to bylo** <podstatné jméno>?
 - Jaké to byly princezny?
 - **Bylo hodně** <přídavné jméno>?
 - Byly hodně mladé?
 - **Pokračuj...**

3. Morfologický generátor

- Šablona definuje odpověď v lematech
 - [jaký] to [být] <podstatné jméno>?
- Ze vstupu uživatele se určí mluvnické kategorie
 - princezen → princezna NN**FP**2 - - - - - A - - - -
 - rod ženský (**F**eminine), číslo množné (**P**lural)
- Morfologický generátor určí slovní formy
 - jaký P4**FP**1 - - - - - - - - - - → jaké
 - být Vp**FP** - - - XR - AA - - - → byly
 - princezna NN**FP**1 - - - - - A - - - - → princezny

Knowledge base

- Chatbot si pamatuje odpovědi uživatele a nikdy se neptá dvakrát na totéž
 - *Uživatel:* Byl jednou jeden král.
 - *Dítě:* Co je to král?
 - *Uživatel:* No to je člověk, který vládne nějaké zemi.
 - (...)
 - *Uživatel:* A tak se Honza rozhodl navštívit krále.
 - *Dítě:* Aha, to je člověk, který vládne nějaké zemi.

Implementace

- implementováno ve frameworku Treex
 - framework pro zpracování přirozeného jazyka
 - tagging, parsing, pojmenované entity...
 - open-source, Perl, vyvíjený na ÚFAL MFF UK
 - čeština, angličtina; základní podpora pro další jazyky
- tagger+generátor MorphoDiTa (Milan Straka)
 - knihovna v C++
 - bindings pro Javu, Python a Perl; web services
 - podporuje libovolný jazyk, jsou-li trénovací data

Demonstrace

Závěr

- Chatbot simulující pohádkychtivé dítě
- Analyzuje vstup od uživatele
 - slovní druhy, mluvnické kategorie
- Odpovídá dle předem daných schémat
 - správné skloňování a časování
- Snadná implementace díky Treexu
 - framework pro zpracování přirozeného jazyka, vyvíjený na ÚFAL MFF UK
- Česká a anglická verze

Možnosti do budoucna

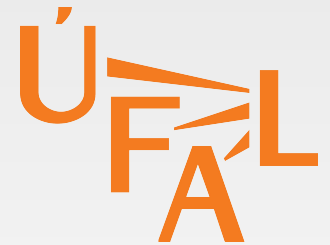
- Hlubší analýza (syntaktická, sémantická?)
- Širší možnosti interakce
- Webové rozhraní
- Řečový dialogový systém
- ... kdokoliv! (GNU GPL v2)

Děkuji za pozornost

Rudolf Rosa
rosa@ufal.mff.cuni.cz

Chatbot Pohádkové dítě
Fairytales Child Chatbot

Univerzita Karlova v Praze
Matematicko-fyzikální fakulta
Ústav formální a aplikované lingvistiky



Stažení chatbota, tato prezentace, a další informace:

<http://ufal.mff.cuni.cz/rudolf-rosa/>