

Návrh softwarového projektu

Motivace: Metody strojového učení našly své místo i v počítačové lingvistice – v oboru, který se zabývá počítačovým zpracováním přirozeného jazyka. Příprava dat pro tyto metody je náročná aktivita ve všech směrech - lidské zdroje, finanční prostředky, časové možnosti.

Využijme proto potenciálu internetových uživatelů, které baví si hrát. Nabídněme jim jazykové hry se slovy, s větami nebo s celými dokumenty, které, aniž by o tom samotní hráči museli vědět, přinesou cenná data pro úlohy počítačového zpracování jazyka.

Cíl projektu: Vytvořit herní portál, který vedle zábavy poskytne cenná data pro úlohy počítačového zpracování přirozeného jazyka. Zaměříme se na sbírání dat hrou pro úlohy, které automatické metody zatím neumí řešit lépe než člověk. Při návrhu her budeme využívat i stávající jazykové zdroje, zejména anotované korpusy.

Příkladem tématu, pro jehož řešení potřebujeme data, je sledování odkazování v dokumentech (důležité např. v systémech strojového překladu). Pro ilustraci uvádíme novinový sportovní článek, ve kterém jsou barevně zvýrazněny příklady slov, která k sobě patří, tj. vzájemně na sebe odkazují. Například o klubu Atletico Madrid se mluví v titulku, první a druhé větě třemi různými slovy – *Atletika, celku, klubu*:

Basile do Atletika

Madrid -

Bývalý trenér argentinského reprezentačního fotbalového výběru Alfio Basile se stal novým kormidelníkem španělského prvoligového celku Atletico Madrid. Basile podepsal smlouvu do června roku 1996 a nahradil na lavičce svého krajana Jorgeho d'Alessandra, kterého prezident klubu Jesus Gil ve středu odvolal jako sedmnáctou "oběť" během svého osmiletého vedení klubu .

Určitě se pokusíme navrhnout a vyzkoušet hru, při které se budou, volně řečeno, „barvit“ slova v dokumentu, jak je naznačeno výše.

Řešení projektu podrobněji:

Přípravná fáze

Seznámit se s počítačovou lingvistikou - úlohy, data, nástroje

Představit témata, pro která budeme sbírat data hrou - jazykové modelování, odkazování v textu, rozpoznávání mluvené řeči, sumarizace textu aj.

Prostudovat spřátelené projekty (viz např. odkaz v části Inspirace)

Analýza

Návrh her (podstata, pravidla, webový design)

Výběr implementačního prostředí her

Výběr vhodné datové struktury jako úložiště „odehraných“ dat (zvážit využití databáze, nebo vlastního datového modelu)

Analýza technologické návaznosti získaných dat na existující zdroje

Implementace

Naprogramování her včetně webového rozhraní

Implementace ukládání a správy získaných dat

Zkušební provoz herního portálu

Doladění systému na základě odezev ze zkušebního provozu

Navrhovaný projekt má vyvážený poměr kreativní a programátorské práce. Začleňuje fáze analýzy, implementace, testování a konečné doladění. K naplnění cíle projektu dojde až po všech těchto fázích.

Inspirace: www.espgame.org

Odhadovaný počet řešitelů: 4

Předpokládaný termín dokončení: září 2008

Kontakt: Mgr. Barbora Vidová Hladká, PhD., ÚFAL MFF UK, e-mail: hladka@ufal.mff.cuni.cz