# SENSEVAL

# SENSEVAL (1)

- Mezinárodní organizace zabývající se hodnocením systémů *Word Sense Disambiguation (WSD)*
- Organizace a řízení hodnocení a související činnosti
- Testování kladů a záporů WSD systémů s ohledem na různost slov, různé aspekty jazyka a různost jazyků

# SENSEVAL (2)

- Prohlubování našich znalostí o lexikální sémantice a mnohoznačnosti
- Řízen malou komisí pod dohledem ACL-SIGLEX (zvláštní zájmová skupina LEXiconu a Assotiation for Computational Linguistics).

# Word Sense Disambiguation (1)

- Zjednoznačnění významu slova
- Problém jak stanovit význam určitého slova v jakémkoliv možném kontextu
- Problém jak formalizovat proces probíhající přirozeně v lidském mozku
- Nejednoznačnost slovních významů je potenciálním zdrojem chyb ve všech aplikacích zpracování jazyka

# Word Sense Disambiguation (2)

- Zažil oživení spolu možností vytvářet velké korpusy
- Díky SENSEVAL se statistické metody, obzvláště techniky strojového učení ukázaly být velmi efektivní
- Podobné značkování části promluvy
  - Zahrnuje označení každého slova speciálním tagem z předdefinované množiny s použitím prvků kontextu a dalších informací

# Word Sense Disambiguation – použití (1)

- Anglické slovo „drug“ ve francouzštině „drogue“ nebo „médicament“
- překladový systém musí rozhodnout o jednoznačném významu slova při každém jeho výskytu
- podobně systémy pro získávání informací mohou chybně vrátit dokumenty týkající se nelegálních narkotik, zatímco dotaz se týkal léků

# Word Sense Disambiguation – použití (2)

- Systémy pro extrakci informací mohou vrátit špatnou odpověď
- Systémy pro převod textu na hlas mohou zaměnit významy slova „zámek“ – budova i předmět k uzamykání

# Word Sense Disambiguation (2)

- V reálných aplikacích bývá WSD většinou plně zabudováno do systému bez možnosti separace
- Např. v systémech pro získávání informací není WSD prováděno explicitně, ale jako vedlejší produkt přiřazování dokumentu k odpovídajícímu dotazu
- Úkolem SENSEVAL je ovšem zaměřit se na samostatné obecně použitelné WSD systémy

# SENSEVAL – historie (1)

- Založena v roce 1997 jako pokračování workshopu SIGLEX-97 - *Značkování s lexikální sémantikou: proč, co a jak?*, pořádaného na konferenci *Aplikace zpracování přirozeného jazyka*
- *Senseval-1* (1998)
  - workshop v Herstmonceux Castle v Anglii 2. až 4. září.
  - úkoly pro angličtinu, francouzštinu a italštinu

# SENSEVAL – historie (2)

- SIGLEX-99
  - Standardizace lexikálních prostředků
  - University of Mariland, červenec 2009

# SENSEVAL-2 (1)

- 2001
- podporován organizacemi EURALEX, ELSNET, EPSRC a ELRA
- ve spojení s ACL-2001
- 5. až 6. července v Toulouse
- Hlavním cílem bylo podpořit účast nových jazyků a vyvinout metodologii pro všeslovné vyhodnocování

# SENSEVAL–2 (2)

- 3 typy úkolů pro 12 jazyků:
  - **„all-words“:** čeština, dánština, angličtina, estonština
  - **Lexikální vzorky:** baskydština, angličtina, italština, japonština, korejština, španělština, švédština
  - **Překlad:** japonština

# SENSEVAL–2 (3)

- **„all-words“** – systémy musí označkovat téměř všechna slova obsažená ve vzorku souvislého textu
- **Lexikální vzorky** – nejprve je ze slovníku pečlivě vybrán vzorek slov. Systém poté musí na krátkém fragmentu textu označkovat určité instance zvolených slov.
- **Překlad** – význam slova je definován vzhledem k odlišnostem v překladu

# SENSEVAL-2 – tabulka odevzdaných projektů

| Language | Task | No. of submissions | No. of teams | IAA | Baseline | Best system |
|---|---|---|---|---|---|---|
| Czech | AW | 1 | 1 | - | - | .94 |
| Basque | LS | 3 | 2 | .75 | .65 | .76 |
| Estonian | AW | 2 | 2 | .72 | .85 | .67 |
| Italian | LS | 2 | 2 | - | - | .39 |
| Korean | LS | 2 | 2 | - | .71 | .74 |
| Spanish | LS | 12 | 5 | .64 | .48 | .65 |
| Swedish | LS | 8 | 5 | .95 | - | .70 |
| Japanese | LS | 7 | 3 | .86 | .72 | .78 |
| Japanese | TL | 9 | 8 | .81 | .37 | .79 |
| English | AW | 21 | 12 | .75 | .57 | .69 |
| English | LS | 26 | 15 | .86 | .51/.16 | .64/.40 |

# SENSEVAL-2 - úkoly

- Skládají se ze 3 typů dat:
  - významový inventář přiřazení slov k významům s eventuálními dalšími informacemi k vysvětlení definici nebo rozlišení významů
  - korpus manuálně značkovaného textu nebo vzorky textu (rozdělen do volitelného cvičného korpusu a testovacího korpusu)
  - volitelná významová seskupení sloužící k omezení významových odlišností ve výsledku

# SENSEVAL-2 – vyhodnocování (1)

- Cílem bylo, aby při nestrojovém značkování nastala shoda v alespoň 90% případů
- Vyhodnocení řízeno centrálně z webu University of Pennsylvania, dodržovalo stejný postup jako v případě SENSEVAL-1
- Pro každý úkol byla data rozdělena do třech částí: zkušební, cvičná a testovací data
- Týmy registrují své systémy, stáhnou data a poté mají 21 dnů na práci s cvičnými daty a 7 dnů na testovací data

# SENSEVAL-2 - vyhodnocování (2)

- Každý tým odevzdá odpovědi na web, kde jsou automaticky vyhodnoceny
- Výsledky SENSEVAL-2 byly zhruba o 14% horší než v SENSEVAL-1
- Mnoho systémů byly pouze vylepšené verze stejných systémů, které se účastnily SENSEVAL-1

# SENSEVAL-2 – plány do budoucna

- Současné nejvýkonnější systémy dosahují svých úspěchů díky kontrolovanému strojovému učení
- Současný výzkum se zaměřuje na to, jak způsoby výběru pro algoritmy strojového učení ovlivňují výkon na různých typech mnohoznačnosti
- Na základě toho, jak snadno nebo obtížně jsou slova zjednoznačněna různými metodami, bychom měli být schopni určit různé typy polysémie

# ACL-02 Workshop (1)

- Word Sense Disambiguation: Současné úspěchy a směry do budoucnosti
- Philadelphia, červenec 2002
- Analyzovat výsledky SENSEVAL-2, plánování SENSEVAL-3
- Hloubková analýza odevzdaných projektů
- Srovnání různých systémů, technik nad různými jazyky

# ACL-02 Workshop (2)

- Srovnání SENSEVAL-1 a SENSEVAL-2
- Co způsobuje, že je pro některá slova snazší najít jednoznačný význam než pro jiná?
- Efektivnost různých korpusů
- Odlišnosti v inventářích významů požadovaných pro různé aplikace

# ACL-02 Workshop (3)

- Zvláštní diskuse věnovaná sémantice předložek
  - Většina z nich je vysoce mnohoznačná
  - Význam je většinou abstraktní
  - Více než u jakékoliv jiné syntaktické kategorie je přesný význam definován kontextem
  - Vítány metody klasifikace předložek a formalizování sémantik

# SENSEVAL – historie (4)

- *Senseval-3* (2004)
  - od března do dubna
  - s navazujícím workshopem v červenci 2004 v Barceloně
  - WSD úkoly pro jazyky baskytština, katalánština, čínština, angličtina, italština, rumunština a španělština
  - obsahoval 14 různých úkolů jak na samotné WSD, tak i na identifikaci sémantických rolí, mnohojazyčné anotace, logické formy, subkategorizační akvizice, atd.
  - více než 55 týmů, přes 160 systémů, 16 úkolů

# SENSEVAL 2007

- Vyhodnocování systémů: 26.2. – 1.4.
- Workshop: 23. – 24. června v Praze spolu s ACL-2007
- Účastníkům budou poskytnuta testovací data
- Podobné úkoly budou testovány nad stejným formátem dat
- 18 úkolů

# SENSEVAL – příklady úkolů

- Klasifikace sémantických vztahů mezi nominály
- Vícejazyčné čínsko-anglické lexikální vzorky
- WSD předložek
- Webové vyhledávání lidí
- Emoční text
- Víceúrovňová sémantická anotace katalánštiny a španělštiny
- Anglické lexikální substituce
- Vyhodnocení znalostních zdrojů o širokém pokrytí

# SEMiSUSANNE

- Sémanticky značkovaný a strukturovaně anotovaný korpus
- Autor: Chris Powell
- Vzniklý sloučením korpusů SUSANNE a SemCor
- Skládá se ze 33 dokumentů společných pro oba korpusy
- Zachovává strukturu SUSANNE, tj. jedno slovo na řádek
- Využívá významy WordNet 1.6

# SUSANNE

- “SUSANNE” = “Surface and underlying structural analysis of natural English”
- 130 000 reprezentativní průřez psanou americkou angličtinou
- Založeno na podmnožině Brown korpusu čítající přes milion slov
- Pokračující projekt CHRISTINE rozšířeno i o mluvenou angličtinu („umm“, „err“)
  - reprezentativní vzorky dnešní mluvené britské angličtiny
- Nejnovější rozšíření LUCY
  - psaná angličtina moderní Británie
  - teenageři, 9-12 leté děti, méně zkušení pisatelé

# SEMiSUSANNE - struktura

- Podobná struktuře SUSANNE
- Osm polí na každé slovo
  - šest z původního korpusu
  - pole indukující samostatné/složené slovo
  - pole pro kódování významu získaného ze SemCor

# SEMiSUSANNE – významové tagy (1)

- Významové pole je složeno ze 2 částí:
  - znak z množiny {n,v,j,r}
    - podstatné jméno, sloveso, přídavné jméno, příslovce nebo část promluvy
  - číselná posloupnost
    - lokalizující element v rámci WordNet databáze
- Např. kódování významu slova „irregularity“:

A01:0030.06 - NN2 irregularities irregularity.Np:s] 0 n475542

# SEMiSUSANNE – významové tagy (2)

- Použité kódování má zajistit
  - rychlý přístup k databázi s použitím WordNet API
  - řeší problém synonym při porovnávání významů s použitím formálních významových klíčů WordNetu

# SEMiSUSANNE – kódování složených slov

- Význam složeniny je přiřazen každé ze složek
  - Opakován na každém odpovídajícím řádku

```
A01:0010.09  -  NP1s    Fulton  Fulton  [Nns.   1       n17954
A01:0010.12  -  NNL1cb  County  county  .Nns]   2       n17954
A01:0010.15  -  JJ      Grand   grand   .       3       n17954
A01:0010.18  -  NN1c    Jury    jury    .Nns:s] 4       n17954
```

# SEMiSUSANNE – kódování samostatných slov

- V poli indikující, zda jde o složené slovo je vždy 0

A01:0010.21 – VVDv said say [Vd.Vd] 0 v682542

# SEMiSUSANNE – funkční slova

- Slova jimž není přiřazen konkrétní význam

A01:0010.06 – AT The the [O[S[Nns:s. - -

# DSO korpus (1)

- 191 lemat, 192 800 instancí
- Inventář: WordNet 1.5
- Zdroje textů: Brown Corpus, Wall Street Journal
- Autoři: Hwee Tou Ng a Hian Beng Lee

# DSO korpus (2)

- Významově značkované výskyty 121 podstatných jmen a 70 sloves
- Nejfrekventovanější nebo víceznačná anglická slova
- 192 800 vět převzatých z korpusu Brown a z Wall Street Journal
- Definice významu jednotlivých výskytů z WordNet 1.5

# DSO korpus (3)

- Obsahuje kompletní výčet definic významů z WordNet 1.5
- Formát věty s významovým tagem pro slovo „action“:

ca01.db #020 `` These >> actions 8 << should serve to protect in fact and in effect the court 's wards from undue costs and its appointed and elected servants from unmeritorious criticisms, " the jury said .

# DSO – definice významu z WordNet1.5

Sense 8 legal action, action, case, lawsuit, suit -- (a judicial proceeding brought by one party against another; "no criminal cases were heard while the judge was ill") => proceeding, legal proceeding, judicial proceeding, proceedings -- (the institution of a legal action) => due process, due process of law -- (the administration of justice according to established rules and principles) => group action -- (action taken by a group of people) => act, human action, human activity -- (something that people do or cause to happen)

# DSO (4)

- Všechny značkované výskyty daného podstatného jména nebo slovesa jsou uloženy společně v jednom souboru
- Každá věta na jednom řádku, definice významu uloženy ve dvou souborech

# Interest korpus (1)

- 1 lemma, 2369 instancí
- Inventář: LDCOE
- Zdroje textů: Wall Street Journal, Rebecca Bruce and Jan Wiebe
- Datový soubor složen z vět obsahující pouze slovo „interest“ nebo „interests“
- Automaticky vyňato z korpusu Penn Treebank Wall Street Journal
- Každá věta v souboru obsahuje jeden významově značkovaný výskyt slova „interest“ („interests“)

# Interest korpus (2)

- Významové značky odpovídají šesti významům slova "interest" definovaných v elektronické verzi první edice Longman's Dictionary of Contemporary English
- Významové značky přidány za slovo
- Např. interest_6/NN
  - význam číslo 6
- Každá věta v souboru je vymezena řádkem obsahující symboly „$$“
- Celkem 2369 vět

# SemCor

- 23 346 lemmat, 234 113 instancí
- Inventář: WordNet 1.6, 1.7, 1.7.1, 2.0
- Zdroje textů: 80% korpus Brown, 20% novela *The Red Badge of Courage*

# Hector

- slovník
- Výzkumný projekt Oxford University Press a DEC
- cca 300 lemmat, 200 000 instancí
- Inventář: Hector
- Zdroj textu: A 20M-word pilot pro British National Corpus

# Open Mind Word Expert

- 230 lemmat, 70 000 instancí včetně duplikátů, počet každým dnem roste
- On-line zdroj, uživatelé ho mohou sami kdykoliv rozšiřovat
- Inventář: WordNet 1.7
- Zdroje textů: Penn treebank, LA Times, aj.

# HKUST-Chinese

- 38 725 vět
- Inventář: Hownet
- Zdroj textu: Sinica corpus

# Švédský korpus

- 179 151 značkovaných instancí
- Inventář: Gothenburg lexical database
- Zdroj textu: The SUC Corpus

# Popisky obrázků

- 2 304 lemmat, 8 816 instancí
- Inventář: WordNet 1.5
- Zdroj textů: Popisky obrázků z obrázkové kolekce

# Line, hard, serve

- 3 slova, více jak 12 000 instancí
  - více než 4000 instancí podstatného jména „line“ značkováno 6 významy wordnetu
  - více než 4000 instancí přídavného jména „hard“ značkováno 3 významy wordnetu
  - více než 4000 instancí slovesa „serve“ značkováno 4 významy wordnetu
- Inventář: WordNet 1.5
- Zdroje textů: Wall Street Journal, American Printing House for the Blind, San Jose Mercury, Leacock, Towell

# Odkazy

- Homepage: www.senseval.org
- www.itri.brighton.ac.uk/events/senseval
- Výsledky SENSEVAL-2: www.itri.brighton.ac.uk/events/senseval/ARCHIVE/RESULTS/senseval-summary.html - unassignable

  http://193.133.140.102/senseval2/Results/all_graphs_files/sheet001.htm