Linguistica Computazionale – Discussione

Traduzione automatica di Titoli di Giornale

Mercoledì 4 Aprile 2007

Cristiano Chesi, chesi@media.unisi.it

Struttura sistema di traduzione A.A. 2005-06

Domande da porsi

- GRUPPO 1 - **raccolta / analisi / traduzione corpus** **(lezione 3, 5)**
  - Tipo di codifica (database, XML o altro standard, informazioni speciali...)
  - Allineamento (cioè traduzione) del corpus multilingue
  - Estrazione lessico, frequenze, concordanze, espressioni idiomatiche...
- GRUPPO 2 - **analisi morfologica** **(lezione 4, 5)**
  - Codifica tratti e categorie
  - Etichettatura PoS (Part-of-Speech, es, [$_{\text{Articolo}}$ la]; [$_{\text{Pronome}}$ la])
  - Algoritmo di analisi (FST, espressioni regolari etc.), adeguatezza del modello
- GRUPPO 3 - **analisi sintattica / parsing / transfer** **(lezione 6, 7, 8)**
  - Selezione PoS corretto
  - Codifica descrizione strutturale
  - Algoritmo di Parsing (eventuale parametrizzazione)
  - Valutazione della reversibilità dell'algoritmo di Parsing
- GRUPPO 4 - **ontologia / rappresentazione della conoscenza / generazione** **(lezione 9, 10, 11)**
  - Rappresentazione dei concetti nel dominio di conoscenza selezionato
  - Uso "creativo" di certe espressioni ("non ha alzato un dito per aiutarmi")
  - Rappresentazioni delle relazioni funzionali da associare alle strutture sintattiche
  - Produzione di testo da strutture sintattiche/rapp. semantiche (scelta approccio: templates, planning)

(con questo colore sono indicate le componenti di pubblico interesse)

Riassunto nodi chiave

- GRUPPO 1 - **raccolta / analisi / traduzione corpus**

1. **Problematiche**

- Tipo di codifica (problemi di efficienza, completezza, scalabilità)
- Allineamento titolo-titolo, chunk-by-chunk, parola-per-parola
- Estrazione lessico, frequenze, concordanze, espressioni idiomatiche...

2. **Strumenti**

- XML, database, treebanks (es. Turin University Treebank)
- analisi statistica (software freeware)
- http://www.ciscl.unisi.it/progetti/news/

3. **Testi di approfondimento**

- Lenci, Montemagni & Pirrelli (2005) *Testo e Computer: Elementi di Linguistica Computazionale*. Carocci, Roma

Riassunto nodi chiave

- **GRUPPO 2 – analisi morfologica / robustezza errori**

1. **Problematiche**

- Codifica tratti e categorie
- Ambiguità PoS
- Iper/Ipo-regolarizzazioni

2. **Strumenti**

- FST (es. PC-Kimmo)
- espressioni regolari (es. GREP)
- dizionari pre-taggati (es. Morph-it)
- http://www.ciscl.unisi.it/childes/

3. **Testi di approfondimento**

- Jurafsky & Martin (2000) *Speech & Language Processing*. Prentice Hall, NJ (Cap. 2)

Riassunto nodi chiave

- **GRUPPO 3 – analisi sintattica/parsing/transfer**

1. **Problematiche**

- Selezione PoS corretto
- Codifica descrizione strutturale
- Valutazione algoritmo di Parsing (ed eventuale parametrizzazione)
- Valutazione della reversibilità dell'algoritmo di Parsing

2. **Strumenti**

- Treebanks (es. TUT, Turin University Treebank)
- FSA/CFG toolkit (es. NLTK, Natural Language Toolkit)
- http://www.ciscl.unisi.it/labN/ (*N*=1,2,3,4,5)

3. **Testi di approfondimento**

- Jurafsky & Martin (2000) *Speech & Language Processing*. Prentice Hall, NJ (Cap. 10, 13)

Riassunto nodi chiave

- **GRUPPO 4 – ontologia/rappresentazione della conoscenza/generazione**

1. **Problematiche**

- Rappresentazione di un dominio specifico di conoscenza (ontologia)
- Approfondimento di un aspetto semantico particolare (es. aspetto-tempo, ambiguità semantica, non-composizionalità)
- Rappresentazioni delle relazioni funzionali da associare alle strutture sintattiche

2. **Strumenti**

- (Multi)WordNet
- http://www.ciscl.unisi.it/progetti/wordnet/

3. **Testi di approfondimento**

- Jurafsky & Martin (2000) *Speech & Language Processing*. Prentice Hall, NJ *Cap. 14, 15(, 16)*

Come avverrà la valutazione del lavoro svolto

- **Chiarezza del lavoro (voto max 28/30)**
  - espressione degli intenti precisa ed esplicita (obiettivi e motivazioni linguistiche e/o computazionali, codifiche esplicite, relazioni con le altre componenti del sistema)
  - eventuali (e adeguati) grafici di supporto (UML, flow-charts...)
- **Analisi delle problematiche (30/30)**
  - Significatività delle problematiche (linguistiche e/o computazionali) affrontate rispetto agli obiettivi proposti
  - individuazione relazioni/pattern significative/i tra i dati
- **Originalità, efficacia delle soluzioni proposte (30 e lode)**
  - approfondimenti/letture indipendenti, riflessioni originali (**non è necessario implementare le soluzioni proposte per prendere la lode**! Ricordatevi che una buona e completa formalizzazione, ad esempio usando i flow-charts, rappresenta l'80% del lavoro di sviluppo!)

Prossima lezione

(Martedì 11 Aprile, ore 16-18, Aula 456, Palazzo S.Niccolò)

- **Approccio sub-simbolico al NLP (e alla MT)**
  - Problemi che si prestano ad un approccio subsimbolico
    - in cosa consiste
    - quando e perché lo si preferisce ad un approccio simbolico
  - Reti neurali
    - neurofisiologia
    - reti di neuroni artificiali
    - apprendimento del passato in inglese
  - NLP con reti neurali: Simple Recurrent Networks
    - memoria a breve termine
    - apprendere incrementalmente il linguaggio
    - usare reti neurali per la MT
    - acquisire proprietà ricorsive con le SRN