Université de Montréal

Information Extraction and Question Answering

Introduction and Context

Guy Lapalme
Université de Montréal
2006

rali

rali

Recherche appliquée en linguistique informatique

Applied Research in Computational Linguistics

français

Research | Technology | Demos | Talks | Collaborators | rali

RALI's personnel includes computer scientists and linguists with considerable experience in natural language processing. It is the largest NLP laboratory in Canada.

Research Projects

- Translation
- Information extraction
- Automatic summarization
- Information retrieval
- Judicial texts processing

System demos

- Translation
  - TransSearch : bilingual concordancer
  - TransType : interactive translation
  - TTPlayer : dynamic TransType trace viewer
- MBOI : Specialized information retrieval
- Reacc : automatic French accentuation
- SILC : language and coding detection
- Lexiqum : Québec text concordancer
- Quantum : Question-answering

Information

- Publications
- Textual Resources

- Contacts
- Members
- Collaborators

- Weekly seminars (mostly in French)

- Courses by professors of the RALI
  - Fall 2006: IFT3330: Intelligence Artificielle
  - Falll 2006: IFT6810: Traitement Statistique des Langues Naturelles
  - Winter 2006: IFT6255 Recherche d'information
  - Winter 2005: IFT6281 : Web sémantique

- *RALI in the press*

A selection of international conferences in NLP

Computer based language processing

Multiple motivations

- Theoretical
  - computational linguistics
- Applicative
  - natural language processing (NLP)
- Cognitive
  - understanding of human performance

rali

Typical Applications of NLP

- Information retrieval
- **Information Extraction**
- **Question-answering**
- Summarization
- Terminological extraction
- Text mining

Information retrieval

- Long tradition of document research
- TREC since 1990
- Revival with the Web
  - comparison of a query with document descriptors
    - criteria
      - statistical
      - morpho-syntactic
      - semantic
    - Performances measured with *precision* and *recall*
- Relatively weak contribution of NLP

rali

Information Extraction

- Find specific information elements in order to fill forms
- Originally defined for MUC (91-97)
  - terrorists acts
  - mergers and acquisitions
- Implemented with rules
  - named entities (NE) finding
  - link between NEs and role identification

rali

Information extraction from conversations (Boufaden 2005)

1 a : Maritime operation centre (ORGANISATION), (INAUDIBLE) hello.

2 b : hi, Mr. Wellington (PERSON), it's captain Mr. VanHorn (PERSON)

3 a : yes.

.....

4 b : ha, Ha, I don't know if I was handled over to you at all, but we've got an overdue boat (VESSEL) on the South Coast of Newfoundland (LOCATION), just in the area quite between Fortune Bay and Trepassey (LOCATION).

5 b : it's on the south east coast of Newfoundland (LOCATION).

.....

6 b : this is been going on for, for 24 hours (TIME) that the case has, or almost anyway, and we had an DFO King Air (AIRCRAFT) up flying (STATUS) this morning (TIME).

7 b : they did (STATUS) a radar search (MEANSOFDETECTION) for us in that area (LOCATION).

8 a : yes.

9 b : and their search (TASK) turned up (STATUS) nothing (RESULT).

10 a : yeah.

Information extraction from conversations (Boufaden 2005) - forms

Search-Mission	
1 ID	*MISSION2*
2 SEARCH-UNIT	*UNIT1, UNIT2*
3 REGION	South East Coast of NewFoundland
4 TYPE	radar search
5 STATUS	being planned

Search-Unit2	
1 ID	*UNIT1*
2 BRAND	King Air
3 ORGANIZATION	DFO
4 DETECTION-EQUIPMENT	radar
5 STATUS	
6 RESULT	nothing
7 DATE	this morning

Question Answering

- Give back an information, not a document reference
- Factual question without dialogue
- Popularized with QA-Track de TREC (1998-)
  - *What ocean did the Titanic sink in?*
  - *Who was the 33rd president of the United States?*
  - *At what speed does the Earth revolve around the sun?*
- EQUER in France (2004)
  - *Où est né Jacques Chirac ?*
  - *Quels sont les signataires du traité de Schengen?*
  - *Qu'est-ce que le SMIC ?*

rali

Question Answering
origins

- deep understanding of situations
- complex semantic representation
- logic reasoning
- Examples :
  - LUNAR, SHURDLU
  - Kalipsos

rali

Question Answering
current approaches

- Question analysis
  - category
  - expected type of answer
  - semantic constraints
  - important words
- Answer searching
  - passage retrieval with important words
  - named entities with important words
  - finding entity satisfying constraints
- Confirmation from external resources ?

rali

Automatic Summarization

- Shortened text keeping only the essential part of a document
- *Indicative* vs *informative*
- Variable length (2% à 50%)
- Reading aid and contextual exploration

Automatic Summarization (Saggion 02)

Designing for human-robot symbiosis

Presents the views on the development of intelligent interactive service robots. The authors have observed that a key research issue in service robotics is the integration of humans into the system. Discusses some of the technologies with particular emphasis on human-robot interaction, and system integration; describes human direct local autonomy (HuDL) in greater detail; and also discusses system integration and intelligent machine architecture (IMA). Gives an example implementation; discusses some issues in software development; and also presents the solution for integration, the IMA. Shows the mobile robot.

Identified Topics: **HuDL - IMA - aid systems - architecture - holonic manufacturing system - human - human-robot interaction - intelligent interactive service robots - intelligent machine architecture - intelligent machine software - interaction - key issue - widely used interaction - novel software architecture - overall interaction - robot - second issue - service - service robots - software - system - Technologies**

Figure 1
Indicative abstract and identified topics for the text “Designing for human-robot symbiosis” D.M. Wilkes et al. Industrial Robot, Vol 26, Issue 1, 1999.

Automatic Summarization
(Saggion 02)

Development of a service robot is an extremely challenging task.

In the IRL, we **are using HuDL** to guide the development of a cooperative service robot team.

IMA is a two-level software architecture for rapidly integrating these elements, for an intelligent machine such as a service robot.

A holonic manufacturing system is a manufacturing system having autonomous but cooperative elements called holons (Koestler, 1971).

Communication between the robot and the human is **a key concern for intelligent service robotics**.

Figure 2
Informative abstract elaborating some topics.

Automatic Summarization

- Mainly statistical approaches
- More *Extraction* than *Abstraction*
- Problems
  - coherence of extracts (anaphora)
  - grammaticality of concatenation
- Summary of many texts on the same subject
- Summary of *non newspaper type* texts
  - scientific articles
  - juridical texts

rali

Term Extraction

- *Term* : linguistic label for a concept
- *Terminology* systemize the knowledge of a domain, only in a hierarchy
- Examples :
  - data, data base
  - cell, blood cell
- often multilingual

rali

Term Extraction - 2

	Known Terms	Unknown Terms
Discovery	*Enrichment*	*Acquisition*
Identification	*Controlled Indexing*	*Free Indexing*

Text mining

- according to Marti Hearst

  the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. A key element is the linking together of the extracted information together to form new facts or new hypotheses

- Variation of *data mining* on textual data (non structured)
- Analogous to a detective work
- Success story:
  - protein interaction detection from cooccurring words in many journal articles

rali

Conclusion

Information Extraction
Question Answering

- specific user needs
- superficial corpus processing
- often fuzzy boundary with other domains (e.g. summarization)
- application oriented approach
- other issues
  - multilingual
  - learning
  - Do we need to understand to answer ?

rali

References

- R. Baeza-Yates, B. Ribeiro-Neto, *Modern Information Retrieval*, Addison-Wesley, 1999.
- M-F. Moens, *Information Extraction: Algorithms and Prospects in a Retrieval Context*, Springer 2006
- E.M. Vorhees, D. K. Harman, *TREC: Experiment and Evaluation in Information Retrieval*, MIT Press, 2005, Chap 10.
- J-L Minel, *Filtrage sémantique (du résumé automatique à la fouille de texte)*, Hermès, 2002.
- C. Jacquemin, *Spotting and discovering terms through Natural Language Processing*, MIT Press, 2001.
- M. Hearst, Text Data Mining, chap 34 in *The Oxford Handbook of Computational Linguistics*, R. Mitkov, ed, 2003.
- Weiss, S.M., Indurkhya, N., Zhang, T., Damerau, F. , *Text Mining Predictive Methods for Analyzing Unstructured Information*, Springer 2005.

rali

Plan for the rest

- Initial work on understanding
- Information extraction
  - named entities
  - superficial grammars
- Question Answering
  - architecture
  - evaluation