International Symposium on Reference Resolution for Natural Language Processing (RRNLP 2002), p. 101-107, Alicante (Spain), June 3-4 2002.

# Automatic Evaluation for Anaphora Resolution in SUPAR system

**Antonio Ferrández; Jesús Peral; Sergio Luján-Mora**

*2002 International Symposium on Reference Resolution for NLP*

UNIVERSITY OF ALICANTE

Departament of Languages and Information Systems

*Group of Language Processing and Information Systems*

## Outline

1. Introduction.
2. The SUPAR system.
3. The automatic evaluation module in SUPAR.
4. Some SUPAR's evaluation results.
5. Conclusions and future works.

## 1. Introduction

- **The comprehension of anaphora is crucial in any application that pretends to deal with NL.**
- **During the last years, there have been many proposals to resolve different kinds of anaphors:**
  - Those that rely on constraints and preference heuristics, Centering Theory, etc.
  - But, there is not a comparative evaluation of all these systems on the same texts and languages since MUC-6, MUC-7.
  - Since then, several efforts to set a common evaluation measures (Barbu and Mitkov, 2001; Byron, 2001) have been carried out.
  - But a comparative evaluation on the same texts is



## 1. Introduction

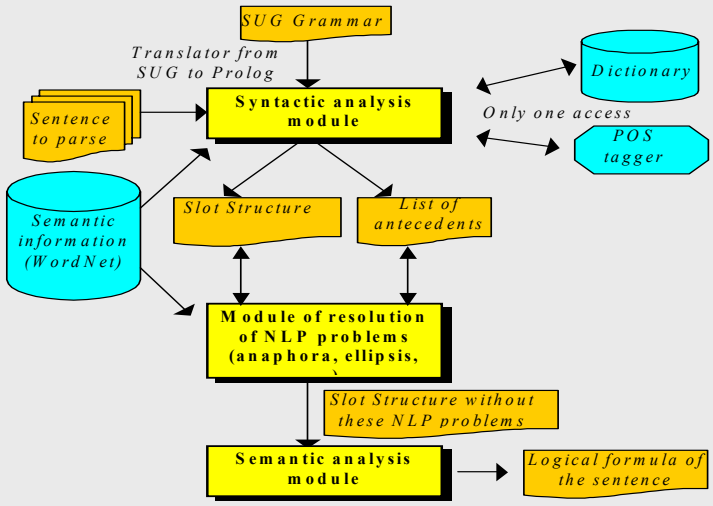
- **Some attempts to carry out a comparative evaluation of anaphora resolution modules:**
  - MUC-6 and MUC-7 co-reference evaluation on the same texts.
  - Independent evaluations on different texts with common evaluation measures (Barbu and Mitkov, 2001; Byron, 2001).
  - Implementing several baselines or well-known strategies on the same language and pre-processing tools.
- **The best approach:**
  - Similar evaluation to MUC or TREC:
    - Same anaphorically tagged texts and languages.
    - Each anaphora resolution module with its own pre-processing tools.

## 2. The SUPAR system

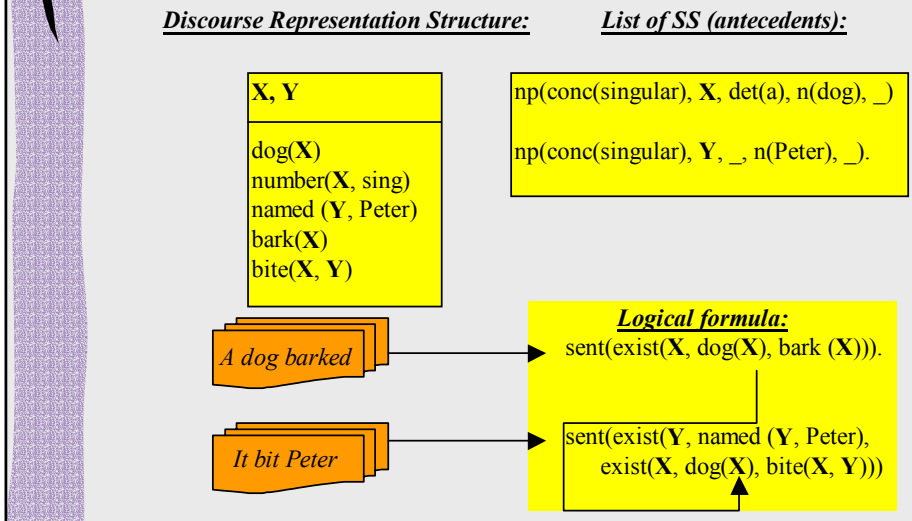
- **The *Slot Unification Parser for Anaphora Resolution, SUPAR*, (Ferrández et al. 1999):**
  - It is a general-purpose NLP system included in a Question Answering system (TREC-9 and TREC-10).
  - It can work on different languages (currently on Spanish or English texts).
  - It carries out either partial or full parsing of the text.
  - It segments the text into sentences and clauses.
  - It can solve pronouns, definite descriptions, zero-



## 2. The SUPAR system. SUPAR architecture



## 2. The SUPAR system. SUPAR anaphora resolution




7 de 19

## 3. The SUPAR evaluation module

- **General description:**
  - It uses as input an anaphorically tagged text.
  - It returns several evaluation measures:
    - The number of sentences and words.
    - The different number of anaphors grouped in categories (reflexive, demonstrative, etc.).
    - The number of candidates before and after restrictions.
    - The number of anaphors resolved just with constraints.
    - The evaluation measures reported in other works (Barbu and Mitkov, 2001; Byron, 2001) such as precision, recall, success rate and critical success rate.

8 de 19

## 3. The SUPAR evaluation module

- **Exit or failure detection:**
  - By comparing only the heads of the solution stored and the head of the solution given by SUPAR.
    - *"Peter saw the boy with the telescope"*.
    - The system returns as solution: "*the boy with the telescope*".
    - The tagged solution is: "*the boy*".
    - It would success with this measure.
  - By comparing the whole solution.
    - It would fail with this measure.



## 3. The SUPAR evaluation module

- **It avoids the error propagation:**
  - Let us suppose that an anaphor is incorrectly resolved.
  - The evaluation module automatically substitutes it in the list of antecedents by the proper solution stored in the tagged text (although it is considered as a failure in the final evaluation).
  - If a following anaphor chooses as its solution the antecedent that is the solution of the previous anaphor, then the second anaphor will not fail in case the first anaphor is incorrectly resolved.

## 3. The SUPAR evaluation module

- **It avoids the error propagation. An example:**
  - By yesterday s close of trading, it was good for a paltry $ 43.5 million. Of course, Mr. Wolf, 48 years old, has some savings.
  - He left his last two jobs at Republic Airlines and Flying Tiger with combined stock-option gains of about $ 22 million, and UAL gave him a $ 15 million bonus when it hired him.



## 3. The SUPAR evaluation module

- **The tagging tool:**
  - It is a semi-automatic anaphor-tagging tool:
    - It can work on text that has been previously POS tagged and segmented into words and sentences.
    - It can work on the output of the SUPAR system:
      - The anaphors detected in the text.
      - Their position in the text: number of sentence and words.
      - The kind of anaphor: e.g. *persRefl* stands for reflexive pronouns.
      - The type of reference: anaphors (<), cataphors (>), exophors (*!*) or any kind of reference (e.g. bound anaphora or references to new objects in discourse: *$*).
      - The position of each possible candidate in the text, a list with those candidates that satisfy constraints, and the final solution.

## 3. The SUPAR evaluation module

- **The tagging tool (cont.):**
  - It can set the co-reference chains, since an anaphor can have as solution another anaphor.
    - In the evaluation module, it would be considered as a right solution whether the system returns as the selected antecedent the other anaphor or its solution.
  - It can tag different kinds of anaphors such as definite descriptions, zero-pronouns, cataphors or exophors.
  - It can anaphorically tag different languages:
    - Currently, we have tagged 921 Spanish pronouns and 1,163 English pronouns.



## 3. The SUPAR evaluation module

- **The automatic evaluation for other systems:**
  - Two different ways of comparison:
    - Each system works with its pre processing tools on the anaphorically tagged texts. They provide the solutions for each anaphor resolved.

| | |
|---|---|
| 3 33 34 persRefl <<br>8 0 1 persIt <<br>... | ANTECEDENTS<br>10 1 4<br>9 20 22<br>...<br>SOLUTION<br>9 20 22 |

    - Each system works with THE SAME pre processing tools.

## 3. The SUPAR evaluation module

- **The automatic evaluation for other systems:**
  - Two different ways of comparison:
    - Each system works with its pre processing tools.
    - Each system works with THE SAME pre processing tools:

| | | |
|---|---|---|
| *<@OOO,1,example of sentence>*<br>*<@CCC>*<br>*<@SNS,suj,comun,person>*<br>*<@NSN>*<br>*Rockwell NNP rockwell*<br>*International NNP international*<br>*Corp. NNP corp.*<br>*<@/NSN>* | *<@GSJ>*<br>*s POS s*<br>*Tulsa NNP tulsa*<br>*unit NN unit*<br>*<@/GSJ>*<br>*<@/SNS,suj,comun,pers>*<br>*<@VBC>*<br>*said VBD said*<br>*<@/VBC>* | *<@/CCC>*<br>*<@CCC>*<br>*<@ANF>*<br>*<@SNS,suj,pronEnglish,,>*<br>*it PPH1R1 it*<br>*<@/SNS,suj,pronEnglish,,>*<br>*<@/ANF>*<br>*…* |



## 3. The SUPAR evaluation module

- **The automatic evaluation for other systems (cont.):**
  - The tagging format proposed in this paper is different from the one used in MUC-6 or MUC-7 although it could be easily exchanged.
  - It will be available in *http://gplsi.dlsi.ua.es/SUPAR*.

## 4. Some SUPAR's evaluation results

- **Anaphora resolution results:**
  - "It" pleonastic pronouns detection: precision of 91% on 970 pronouns of the TREC Federal Register collection.
  - Spanish zero-pronouns, personal or demonstrative pronouns on texts of different genres (newspapers, technical manuals, novels, etc.): 921 / 1,144 = 81%.
  - English pronouns: 835 / 1,163 = 74%.



## 4. Some SUPAR's evaluation results

- **Efficiency of SUPAR:**
  - 887 randomly selected documents of the TREC collections: the Los Angeles Times (LAT) and the Foreign Broadcast Information Service (FBIS):
    - Parsing time: up to 2,001 words per second.
    - Global SUPAR speed up to 256 words per second.
    - Anaphora resolution module takes about 89% of the total running time (216 reflexive pronouns, 8,722 personal and demonstrative pronouns, 396,977 candidates, 17.8 candidates per non reflexive pronoun after constraints).
    - Pentium III, 1000 GHz, 128 Mb RAM.

## 5. Conclusions and future works

- We have described the evaluation module that has been included in the SUPAR system.
- It automatically evaluates different kinds of anaphors: pronouns, zero-pronouns, and definite descriptions.
- It can work on texts in different languages.
- It has also been presented a tool that facilitates the anaphorical annotation of texts.
- It will allow the comparison with other systems working whether their own pre-processing tools or not.