
Semantic categories I: City, Year, Person, University

Semantic categories II: Place, Time, Person, Organization

Semantic categories III: Geoadmin location, Date, Famous writer, Cultural premise/facility

From Santos (2007)
Evaluation pitfalls because of “same” measure

• the best system in MUC attained F-measure greater than 95%

-> so, if best scores in HAREM had F-measure of 70%, Portuguese lags behind...

CONLL, Sang (2002)

• Several problems:
  – the evaluation measures
  – the task definition

Study at the Temple University’s Graduate School of Business

MUC-7, Chinchor (1997)
Evaluation measures used in MUC and CoNLL

MUC: Given a set of semantically defined categories expressed as proper names in English
- universe is: number of correct NEs in the collection
- recall: number of correct NEs returned by the system/number of correct NEs

CoNLL\textsuperscript{fict}: Given a set of words, marked as initiating or continuing a NE of three kinds (+MISC)
- universe: number of words belonging to NEs
- recall: number of words correctly marked by the system/number of words
Detailed example, MUC vs. CoNLL vs. HAREM

U.N. official Ekeus heads for Baghdad 1:30 pm Chicago time.


[ORG U.N.] official [PER Ekeus] heads for [LOC Baghdad] [TIME 1:30 p.m. [LOC Chicago] time]. (MUC)

[PER U.N. official Ekeus] heads for [LOC Baghdad] [TIME 1:30 p.m. Chicago time]. (HAREM)

From Santos (2007)
Detailed example, MUC vs. CoNLL vs. HAREM

He gave Mary Jane Eyre last Christmas at the Kennedys.

• He gave [PER Mary] [MISC Jane Eyre] last [MISC Christmas] at the [PER Kennedys]. (CoNLL)

• He gave [PER Mary] Jane Eyre last Christmas at the [PER Kennedys]. (MUC)

• He gave [PER Mary] [OBRA Jane Eyre] last [TIME Christmas] at the [LOC Kennedys]. (HAREM)
Task definition

MUC: Given a set of semantically defined categories expressed as proper names (in English) (or number or temporal expressions), mark their occurrence in text
• correct or incorrect

HAREM: Given all proper names (in Portuguese) (or numerical expressions), assign their correct semantic interpretation in context
• partially correct
• alternative interpretations
Summing up

• There are several choices and decisions when defining precisely a task for which an evaluation is conducted

• Even if, for the final ranking of systems, the same kind of measures are used, one cannot compare results of distinct evaluations
  – if basic assumptions are different
  – if the concrete way of measuring is different

From Santos (2007)