LXGram in the Shared Task "Comparing Semantic Representations"

STEP 2008

António Branco[1] Francisco Costa[1]

[1]Universidade de Lisboa

September 20, 2008

Outline

System Description

- LXGram
- Hand-crafted precision grammar for Portuguese
- Deep linguistic processing
- HPSG framework [Pollard and Sag, 1994]
- MRS format of semantic representations [Copestake et al., 2005]
- Developed in the LKB [Copestake, 2002]
- Size of the grammar:
  - 24484 lines of code (including comments)
  - 53 syntax rules
  - 40 lexical rules
  - 3154 total types
  - 414 types for lexical items
  - 2718 hand-built lexical entries

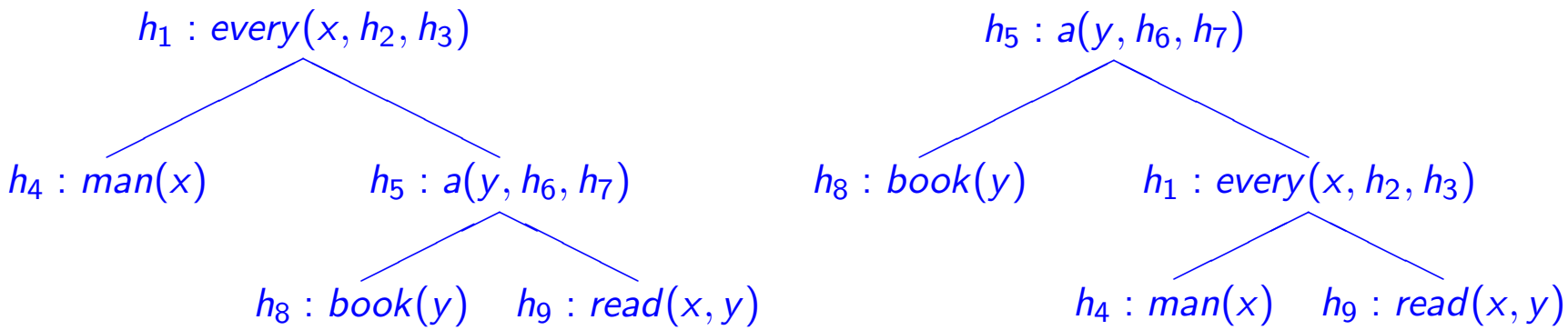
Semantic Formalism

- Minimal Recursion Semantics (MRS)

Semantic Formalism

- Minimal Recursion Semantics (MRS)
- Underspecification of scope

$every(x, man(x), a(y, book(y), read(x, y)))$
$a(y, book(y), every(x, man(x), read(x, y)))$



$\langle h_0, \{\ h_1 : every(x, h_2, h_3),\quad h_4 : man(x),$
$\quad h_5 : a(y, h_6, h_7),\ \ h_8 : book(y),\ \ h_9 : read(x, y)\},$
$\{\ h_0 =_q h_9, h_2 =_q h_4, h_6 =_q h_8\}\rangle$

Semantic Formalism

- Minimal Recursion Semantics (MRS)
- Underspecification of scope
- Easy to represent with feature structures

$$\left[\begin{array}{ll}
\text{TOP} & \boxed{h_0} \\
\text{RELS} & \left\langle \left[\begin{array}{ll} \text{LBL} & \boxed{h1} \\ \text{PRED} & \textit{every} \\ \text{ARG0} & \boxed{x} \\ \text{RSTR} & \boxed{h2} \\ \text{BODY} & \boxed{h3} \end{array}\right], \left[\begin{array}{ll} \text{LBL} & \boxed{h4} \\ \text{PRED} & \textit{man} \\ \text{ARG0} & \boxed{x} \end{array}\right], \left[\begin{array}{ll} \text{LBL} & \boxed{h5} \\ \text{PRED} & \textit{a} \\ \text{ARG0} & \boxed{y} \\ \text{RSTR} & \boxed{h6} \\ \text{BODY} & \boxed{h7} \end{array}\right], \left[\begin{array}{ll} \text{LBL} & \boxed{h8} \\ \text{PRED} & \textit{book} \\ \text{ARG0} & \boxed{y} \end{array}\right], \left[\begin{array}{ll} \text{LBL} & \boxed{h9} \\ \text{ARG0} & \boxed{x} \\ \text{ARG1} & \boxed{y} \end{array}\right] \right\rangle \\
\text{HCONS} & \left\langle \left[\begin{array}{ll} \text{HARG} & \boxed{h0} \\ \text{LARG} & \boxed{h9} \end{array}\right], \left[\begin{array}{ll} \text{HARG} & \boxed{h2} \\ \text{LARG} & \boxed{h4} \end{array}\right], \left[\begin{array}{ll} \text{HARG} & \boxed{h6} \\ \text{LARG} & \boxed{h8} \end{array}\right] \right\rangle
\end{array}\right]$$

$$\begin{aligned}
\langle h_0, \{\ & h_1 : \textit{every}(x, h_2, h_3), \quad h_4 : \textit{man}(x), \\
& h_5 : a(y, h_6, h_7), \quad h_8 : \textit{book}(y), \quad h_9 : \textit{read}(x, y)\}, \\
& \{\ h_0 =_q h_9, h_2 =_q h_4, h_6 =_q h_8\}\rangle
\end{aligned}$$

Semantic Formalism

- Conjunction is represented by identity of handles

$$fat(x) \wedge black(x) \wedge cat(x)$$

$$\left[\ \text{RELS}\left\langle \begin{bmatrix}\text{LBL} & \boxed{h1}\\ \text{PRED} & fat\\ \text{ARG0} & \boxed{x}\end{bmatrix}, \begin{bmatrix}\text{LBL} & \boxed{h1}\\ \text{PRED} & black\\ \text{ARG0} & \boxed{x}\end{bmatrix}, \begin{bmatrix}\text{LBL} & \boxed{h1}\\ \text{PRED} & cat\\ \text{ARG0} & \boxed{x}\end{bmatrix}\right\rangle\ \right]$$

- No need to compute commutativity or associativity of conjunction when working with different languages

$$fat(x) \wedge black(x) \wedge cat(x)$$
$$gato(x) \wedge preto(x) \wedge gordo(x)$$

cat black fat

Semantic Formalism

- Conjunction is represented by identity of handles
- Neo-davidsonian representation of events

"arrives today"

$$arrive(e_1, x) \wedge today(e_2, e_1)$$

$$\left[\ \text{RELS}\ \left\langle \begin{bmatrix} \text{LBL} & \boxed{h1} \\ \text{PRED} & arrive \\ \text{ARG}0 & \boxed{e2} \\ \text{ARG}1 & \boxed{x3} \end{bmatrix}, \begin{bmatrix} \text{LBL} & \boxed{h1} \\ \text{PRED} & today \\ \text{ARG}0 & \boxed{e4} \\ \text{ARG}1 & \boxed{e2} \end{bmatrix} \right\rangle \right]$$

Semantic Formalism

- Conjunction is represented by identity of handles
- Neo-davidsonian representation of events
- Generalized quantifiers

"John sees it"

$$\begin{bmatrix} \text{TOP} & \boxed{h0} \\ \text{RELS} & \left\langle \begin{bmatrix} \text{LBL} & \boxed{h1} \\ \text{PRED} & \textit{proper\_q\_rel} \\ \text{ARG0} & \boxed{x2} \\ \text{RSTR} & \boxed{h3} \\ \text{BODY} & \boxed{h4} \end{bmatrix}, \begin{bmatrix} \text{LBL} & \boxed{h5} \\ \text{PRED} & \textit{named\_rel} \\ \text{ARG0} & \boxed{x2} \\ \text{CARG} & \textit{John} \end{bmatrix}, \begin{bmatrix} \text{LBL} & \boxed{h6} \\ \text{PRED} & \textit{see} \\ \text{ARG0} & \boxed{e7} \\ \text{ARG1} & \boxed{x2} \\ \text{ARG2} & \boxed{x8} \end{bmatrix}, \begin{bmatrix} \text{LBL} & \boxed{h9} \\ \text{PRED} & \textit{pronoun\_q\_rel} \\ \text{ARG0} & \boxed{x8} \\ \text{RSTR} & \boxed{h10} \\ \text{BODY} & \boxed{h11} \end{bmatrix}, \begin{bmatrix} \text{LBL} & \boxed{h12} \\ \text{PRED} & \textit{pronoun\_n\_rel} \\ \text{ARG0} & \boxed{x8} \end{bmatrix} \right\rangle \\ \text{HCONS} & \left\langle \begin{bmatrix} \text{HARG} & \boxed{h0} \\ \text{LARG} & \boxed{h6} \end{bmatrix}, \begin{bmatrix} \text{HARG} & \boxed{h3} \\ \text{LARG} & \boxed{h5} \end{bmatrix}, \begin{bmatrix} \text{HARG} & \boxed{h10} \\ \text{LARG} & \boxed{h12} \end{bmatrix} \right\rangle \end{bmatrix}$$

Semantic Formalism

- Conjunction is represented by identity of handles
- Neo-davidsonian representation of events
- Generalized quantifiers
- Naming convention for predicate names

- Initial underscore for relations that correspond to lexical items
- Lemma
- Part of speech
- Optional sense field
- *_rel* suffix

_house_n_rel

_different_a_from_rel

Semantics Produced by LXGram

- No encoding of thematic roles
- No anaphora resolution
- No general word sense disambiguation

- Word sense can be disambiguated in some cases if it has morpho-syntactic correlates and they are present in a given input sentence.
  E.g.: Pt nabo = En turnip, incompetent person
  2 lexical entries (different predicate names), sencond one can inflect for gender
  If "naba" (fem. form) occurs in the input, it can be disambiguated
  All occurrences of "nabo" will make the grammar posit multiple analyses

Semantics Produced by LXGram

- No encoding of thematic roles
- No anaphora resolution
- No general word sense disambiguation
- All morphological information is encoded directly under the features of events and referential indices (e-type variables)

- tense, aspect, mood

$$\begin{bmatrix} \text{LBL} & \boxed{h1} \\ \text{PRED} & \textit{\_arrive\_v\_rel} \\ \text{ARG0} & \boxed{e2}\begin{bmatrix} \text{E} & \begin{bmatrix} \text{TENSE} & \textit{tense} \\ \text{ASPECT} & \textit{aspect} \\ \text{MOOD} & \textit{mood} \end{bmatrix} \end{bmatrix} \\ \text{ARG1} & \boxed{x3} \end{bmatrix}$$

Semantics Produced by LXGram

- No encoding of thematic roles
- No anaphora resolution
- No general word sense disambiguation
- All morphological information is encoded directly under the features of events and referential indices (e-type variables)

- tense, aspect, mood
- person, number, gender

$$\begin{bmatrix} \text{LBL} & \boxed{h1} \\ \text{PRED} & \textit{\_house\_n\_rel} \\ \text{ARG0} & \boxed{x2} \begin{bmatrix} \text{PNG} & \begin{bmatrix} \text{PERSON} & \textit{person} \\ \text{NUMBER} & \textit{number} \\ \text{GENDER} & \textit{gender} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Sample Text

- Text 4
- Adapted from newspaper text
- 5 sentences, 10–26 words long, avg. 18.6 words
  - English: 11–29 words, avg. 21.8 words
- Interesting phenomena:
  - relative clauses
  - noun ellipsis
  - null subjects (in the Portuguese version)
  - coordination
  - intersective and non-intersective adjectives
  - relational nouns with realized and missing complements
  - ...

Initial Results

- 7 texts, 30 sentences total
- Shared Task data translated into Portuguese by the authors
  - Translations as literal as possible
- Sentence length: 4–49 words, avg. 19.1 words
  - (English: 4–47 words, avg. 18.5 words)
- Initial results
  - 0% coverage on the other 6 texts of the shared task
  - 16.7% coverage overall
- Causes:
  - unknown words
  - unimplemented phenomena
- Lexicon and grammar expansion

Grammar Expansion

- Added 97 lexical entries
  - 10 new lexical types
- Added 3 new syntax rules
  - NP apposition
  - an idiomatic type of temporal expressions
  - subject clitics
- Some constructions deliberately not implemented
  - Wh- questions

Final Results

- Exhaustive search
- Best parse chosen manually
- 20 sentences parsed (66.7%)
- 1–3162 analyses per sentence, avg. 245
- 5KB–1.1 GB memory per sentence
- 253K unification operations total

Discussion of the Results

- No standard way of representing intensionality in MRS

"Researchers have been looking for **other cancers that may be caused by viruses**." (Text 2)

Discussion of the Results

- No standard way of representing intensionality in MRS
- PPs attaching higher than determiners

"federal government interest and tax incentives." (Text7)

Pt	o	interesse	e	os	benefícios	fiscais	[do	governo federal]
gloss	the	interest	and	the	incentives	tax	of the	government federal

Discussion of the Results

- No standard way of representing intensionality in MRS
- PPs attaching higher than determiners
- Restrictive and non-restrictive relative clauses

"We would like our school to work similarly to the French ones, [which live from donations (. . .) and even from the raffles [that children sell at school.]]" (Text 4)

Discussion of the Results

- No standard way of representing intensionality in MRS
- PPs attaching higher than determiners
- Restrictive and non-restrictive relative clauses
- Phenomena beyond the scope of the grammar

"a crewman (. . .) yelled into the phone, "I have a problem here. I am not ready yet." (Text 5)

References

Copestake, A., Flickinger, D., Sag, I. A., and Pollard, C., 2005. Minimal Recursion Semantics: An introduction. *Journal of Research on Language and Computation*, 3(2–3):281–332.

Copestake, A., 2002. *Implementing Typed Feature Structure Grammars*. CSLI Publications, Stanford.

Pollard, C. and Sag, I., 1994. *Head-Driven Phrase Structure Grammar*. Chicago University Press and CSLI Publications, Stanford.