The Typology of Ellipsis: A Corpus for Linguistic Analysis and Machine Learning Applications

Damir Cavar (Indiana University, NLP-Lab)

Ludovic Vetea Mompelat (University of Miami)

Muhammad S. Abdo (Indiana University, NLP-Lab)

+ the NLP-Lab Team

SIGTYP 24, March 2024

Agenda

- Ellipsis Constructions and Syntax
- The Hoosier Ellipsis Corpus
- Evaluations and Results
- Discussion of ML Experiments

Ellipsis Constructions

- Common phenomena like gapping, sluicing, forward or backward conjunction reduction
  - Lexical elements are elided under certain conditions
  - Native speakers have no cognitive issues processing and understanding ellipsis constructions
- Examples...

Ellipsis Constructions

Forward Conjunction Reduction (Across-the-board movement):

- *My sister lives in Utrecht and ___ works in Amsterdam.*

  → *My sister lives in Utrecht and (my sister/she) works in Amsterdam.*

Gapping

- *Paul and John were watching the news, and Mary ___ a movie.*

  → *Paul and John were watching the news, and Mary (was watching) a movie.*

- *Will Jimmy greet Jill first, or ___ Jill ___ Jimmy ___ ?*

  → *Will Jimmy greet Jill first, or (will) Jill (greet) Jimmy (first) ?*

Ellipsis Constructions

- **Discourse Licensed Ellipsis:**
  - A: *Who wants to marry whom?*

    B: *Susan ___ Larry.*

    → *Susan wants to marry Larry.*
- **Semantic Issues:**
  - *John* **[AGENT]** *drove to Wisconsin and ___* **[PATIENT]** *was arrested in Illinois.*
  - *Peter stole a book and John ___ kisses from Mary.*

    → *Peter stole a book and John (stole) kisses from Mary.*

Ellipsis Constructions

- Publicly available datasets:
  - Sluicing corpus for English (Anand et al. 2021)
  - VP-ellipsis corpus for English (Bos & Spenader, 2011; Goldberg & Stubbs 2020)
  - ELLie corpus for English (Testa et al. 2023)
- Small datasets
- Limited to English and a few common languages
- Limited to specific ellipsis phenomena (gapping, sluicing, VP-ellipsis, …)

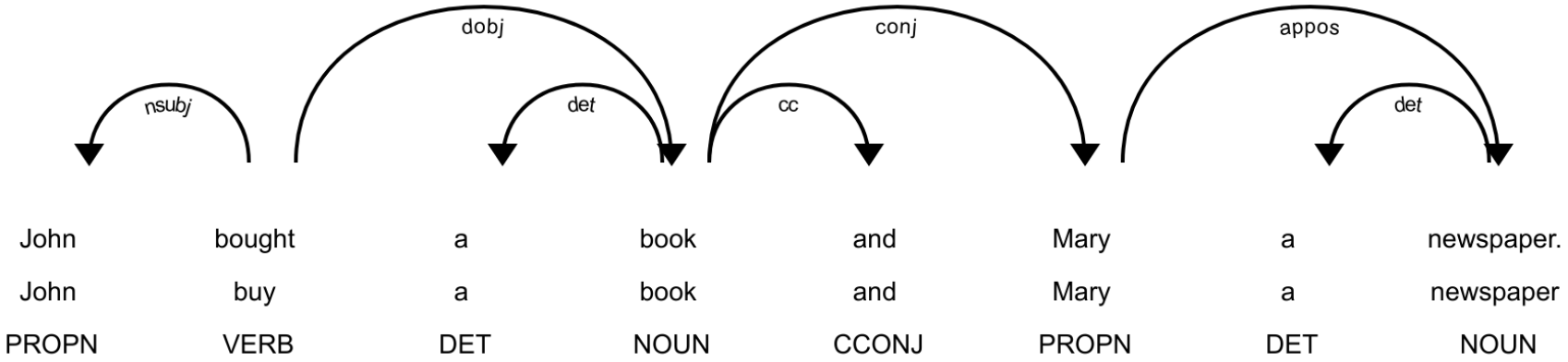
Ellipsis Constructions

- Lack of a cross-linguistic typological overview of ellipsis types
- Explanatory theoretical analysis of ellipsis constructions
- Frameworks like Dependency Grammar, Lexical-functional Grammar, and even Generative frameworks like Minimalist Program do not provide descriptive or explanatory means

Ellipsis Constructions

- Current State of the Art (SOTA) Natural Language Processing-pipelines and parsers perform poorly (or not at all)
- Tested SOTA parsers:
  - Stanford CoreNLP
  - Stanford Stanza (V 1.6) (Dependency & Constituent Parser)
  - Berkley Neural Parser (benepar)
  - SpaCy 3.6
  - XLE (Web-XLE, Lexical-functional Grammar Parser)
- All parsers fail with Ellipsis (and other constructions) → not useful for downstream NLP tasks (e.g., relation extraction)

Dependency Parsers: SpaCy 3.6



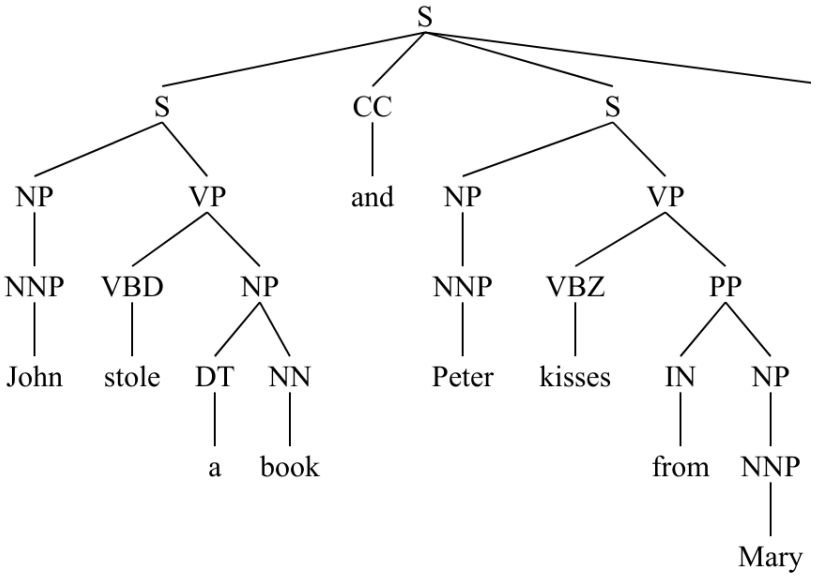
Resulting assumption:

John bought: (a book and Mary) (local coordination of two noun phrases); “a newspaper” is assumed to be a modifier or specifier of “Mary”

Constituent Parsers

Berkley Neural Parser Head

Noun of the object (kisses) is assumed to be the predicate head of the second conjunct.



Computational Tests

- **Cloze test:**
  - Used in Machine Learning – Marked Word Prediction in BERT (LM)
    - *The house ___ I was born.* (a. *where* , b. *which*)
  - Next word prediction as in Large Language Models (LLMs)
- **Tasks:**
  - Classification of sentences / utterances: Does it contain ellipsis or not?
  - Detection of locus of ellipsis: indicate the space
  - Guess of the missing words: fill in the missing words

Experiments

18 Languages with varying number of examples.

- **Largest**: Arabic, Mandarin Chinese, Croatian, English, German, Gujarati, Hindi, Japanese, Kumaoni, Korean, Navajo, Norwegian, Polish, Russian, Spanish, Swedish, Ukranian.
- **In prep**: Bengal, Hebrew, Kanada, Tamil, Telugu
- **Tested:** English, Arabic, Spanish, Russian

- **Picked:**
  - 500 target sentences
  - 1000 distractors
  - For tasks 2 & 3: only examples with ellipsis are used.
- **Algorithms:**
  - Logistic Regression
  - BERT/RoBERTa-based Deep Learning model
  - GPT-4 Large Language Model (ChatGPT), Falcon2, Llama2, etc.

Corpus Access

- In the next days: See NLP-Lab page
  - https://nlp-lab.org/ellipsis/

- Link to GitHub, allowing for collaboration and contribution.
  - https://github.com/dcavar/hoosierellipsiscorpus

Experiments

- **For Arabic:**
  - We utilized GPT-4 (no other LLM was capable of processing Arabic)
    - Missing useful BERT-type LM for Arabic, we need to train one
  - Task 1: 0-shot classification
    - Baseline: Logistic Regression **83%**
    - GPT-4 : Precision 0.56, Recall 0.18, Accuracy 72%
  - Task 3: 0-shot word filling
    - GPT-4 : Accuracy ~80%

Experiments

- **For English:**
  - We utilized GPT-4 (other LLMs failed to provide significant results)
  - Task 1: 0-shot classification
    - Baseline: Logistic Regression 74%
    - GPT-4 : Precision 0.756, Recall 0.599, Accuracy 66,8%
  - Task 3: 0-shot word filling
    - GPT-4 : Accuracy 25%

Baseline Classifier Task 1

- LR → supervised training
  - Training: 1,600 (50% ellipsis constructions)
- Accuracy: 74%
- Can be improved with a few more features, incl. unsupervised feature generation.

LLM Classifiers Task 1

- 0-shot classification: “Does this sentence contain ellipsis?”
- LLMs:
  - GPT 3.5
  - GPT 4
  - Llama2
  - Zephyr
  - Ongoing: Claude 3

LLM Classifiers Task 1

Model	Precision	Recall	F1-Score	Accuracy
GPT 3.5	0.33	0.44	0.38	0.35
GPT 4	0.55	0.67	0.60	0.60
Llama2	0.40	0.67	0.50	0.40
Zephyr	0.25	0.11	0.15	0.42

Preliminary Results

- Supervised ML/NLP methods outperform all LLMs on 0-shot
- GPT with default temperature (0.7)
    - Randomizes 20% of the output decisions, i.e. for 20% of repeated tasks with the same data the classifier will be switched.
- GPT with temperature set to 0
    - No random decisions → deterministic, but:
    - Drop of accuracy by 10% over sample data

LLM Position Guesser Task 2

GPT 3.5	GPT 4	Llama2
Accuracy 0.05	Accuracy 0.15	Accuracy 0.00

- Issues:
  - Prompt engineering and instructions
  - Evaluation and position matching

LLM Missing Word Guesser Task 3

GPT 3.5	GPT 4	Llama2
Accuracy 0.00	Accuracy 0.25	Accuracy 0.00

- Issues:
  - More experiments with prompts.
  - String matching evaluation.

Experiments

English in comparison:

- Task 1:
  - Logistic Regression (baseline): accuracy 74%
  - BERT-based Transformer: accuracy 94%
  - GPT-3.5: accuracy: 35%
  - GPT-4: accuracy: 60%

BERT/Transformer > Logistic Regression > GPT-4

Conclusion

- Problems with "invisible words" in all parsers and LLMs
  - Parsers perform without a problem with "ellipsis undone"

- The problem is:
  - Theoretical – Dependency Grammar, Lexical-functional Grammar, etc.
  - Data-based – missing corpora with annotated ellipsis constructions
  - Computational – LLMs predict next words, and not next missing words (while BERT is trained on masked words)

Thank you ~~for listening~~!