Ellipsis in Arabic: Using Machine Learning to Detect and Predict Elided Words

Muhammad Abdo, Damir Cavar, Billy Dickson (+ the NLP-Lab Team)

Indiana University at Bloomington, Natural Language Processing Lab

The 37th Annual Symposium on Arabic Linguistics

February 2024

Agenda

- Ellipsis Constructions and Syntax
- The Hoosier Ellipsis Corpus
- Arabic Sub-Corpus and Results
- Machine Learning Experiments

SECTION 1

Introduction and Motivation

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 drove to Wisconsin and ___ was arrested in Illinois.*

- *Peter stole a book and John ___ kisses from Mary.*

Ellipsis Constructions

- Publicly available datasets:
  - Sluicing corpus for English
  - VP-ellipsis corpus for English
  - ELLies corpus for English
- Small datasets
- Limited to English and a few common languages
- Limited to specific ellipsis phenomena (gapping, sluicing, VP-ellipsis, …)

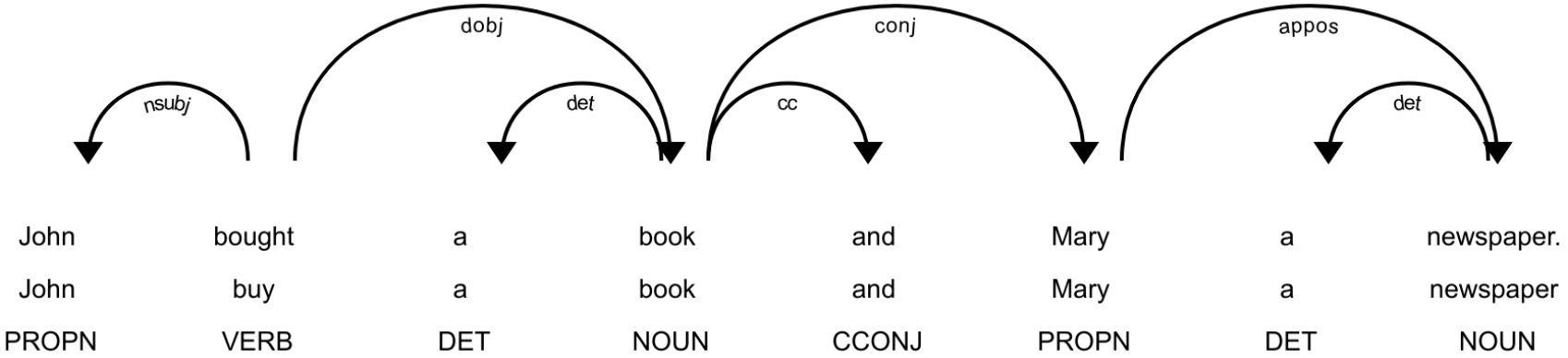
Ellipsis Constructions

- 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

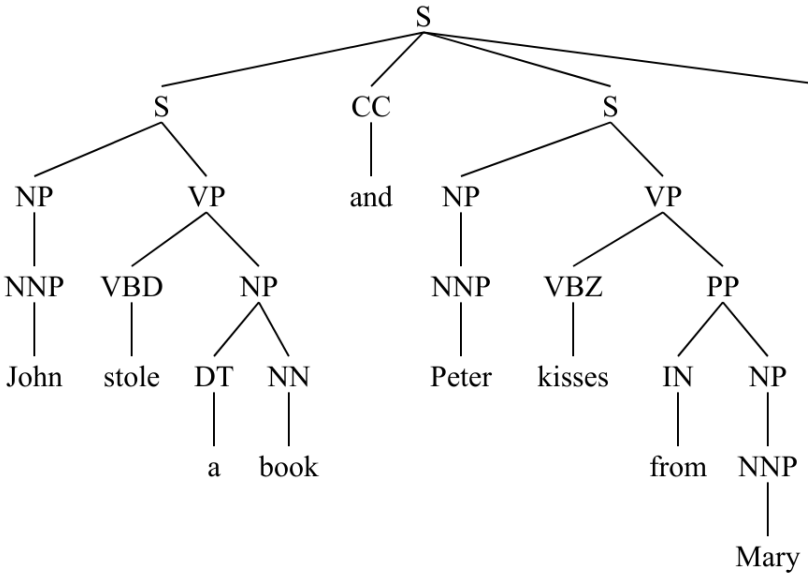
- Current State of the Art (SOTA) Natural Language Processing-pipelines and parsers perform poorly (or not at all)
- Tested SOTA parsers:
  - Stanford CoreNLP
  - Stanford Stanza (V 1.6) (Dependency & Constituent Parser)
  - Berkley Neural Parser (benepar)
  - SpaCy 3.6
  - XLE (Web-XLE, Lexical-functional 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:**
  1. Used in Machine Learning – Marked Word Prediction in BERT (LM)
     - *The house ___ I was born.* (a. *where* , b. *which*)
  2. Next word prediction as in Large Language Models (LLMs)
- **Tasks:**
  1. Classification of sentences / utterances: Does it contain ellipsis or not?
  2. Detection of locus of ellipsis: indicate the space
  3. Guess of the missing words: fill in the missing words

Experiments

- 18 Languages with varying number of examples.
  - Largest: Arabic, English, Spanish
  - Included: Navajo, Gujarati, Croatian, Russian, Polish, Ukranian, Chinese, ...
- Picked:
  - 500 target sentences
  - 1000 distractors
  - For task 2 & 3: only eamples with ellipsis are used.
- Algorithms:
  - Logistic Regression
  - BERT/RoBERTa-based Deep Learning model
  - GPT-4 Large Language Model (ChatGPT), Falcon2, Llama2, etc.

SECTION 3

Data Collection

Building the MSA Corpus
[SketchEngine: ArTenTen]

Sample Corpus Query Language Patterns	Ellipsis Type	Sample Output
[word = "عن	من| في| إلى"] [word = "وأخرى| وآخر"]	Nominal Ellipsis	ورواية عن السودان ورواية أخرى عن الكويت
[word="؟"] [tag="noun"] ([word="!"] | [lemma="."])	Fragment Answer	ما هو أكثر شيء يسعدك في هذه الدنيا؟ المال هو أكثر شيء يسعدني في هذه الدنيا.
[word = "و"] [tag = "noun"] [word=أيضا |كذلك"] [word="."]	Stripping	يتأثر النمو بالوراثة بشكل كبير والتغذية كذلك تتأثر بشكل كبير.
[word = "ولكن"] [word = "أين |الماذا |متى |من |كيف"] [word="؟"]	Sluicing	لدينا قناعة بأن الكل سيعترف بفلسطين، ولكن متى سيعترف الكل بفلسطين؟ سوف تحل المشكلة بنفسها، ولكن كيف ستحل المشكلة بنفسها؟
[tag = "pron|noun"] [tag = "verb"] [tag = "noun"] [word = "و"] [tag = "pron|noun"] [tag = "noun"]	Verbal Ellipsis	هي تهتم باللاوعي وهو يهتم بالعقل
[tag = "pron"] [tag="verb"] [tag="noun"] [] [word="و"] [tag="pron"] [tag = "noun"]	Gapping	أنا نمت فوق السرير وهو نام على الأرض

Building the Egyptian Arabic Corpus
[X AKA Twitter]

Manually searching Twitter for the same patterns of MSA but in Egyptian Arabic.

جميل شكل الناس و هي بتحقق أحلامها..
- يارب و إحنا كمان

Translate post

9:59 AM · Feb 21, 2024 · **1,513** Views

Translation

People look so nice when they achieve their dreams! God! I hope we also ~~achieve our dreams~~.

Data Structure

Sentence with ellipsis expressed by 4 underscores.	هتقضوا رأس السنة فين ؟ ____ في البيت
Separator: 4 dashes	----
Full sentence without ellipsis	هتقضوا رأس السنة فين ؟ هنقضي رأس السنة في البيت
Sentence Source	#Source:Twitter

Translation

Where are you going to spend New Year's Eve? ~~We are going to spend New Year's Eve~~ at home.

Corpus Access

- In the next days: See NLP-Lab page
  - https://nlp-lab.org/ellipsis/
- Link to GitHub, allowing for collaboration and contribution.

SECTION 3

Ellipsis in Modern Standard and Egyptian Arabic

Nominal Ellipsis

فدراسة تتحدث عن الفوائد و~~دراسة~~ أخرى عن المضار (ArTenTen).

A study discusses the benefits, and another ~~study~~ discusses the harms.

الحكومة هي الي بتشتري ~~دولارات~~ من السوق السودا (Twitter)

It is the government which buys ~~dollars~~ from the black market.'

Ellipsis within the noun phrase or when the whole noun phrase is elided; It is also characterized by preserving the syntactic properties of agreement.

(Zdravkovska-Adamova, 2017; Merchant, 2018; Saab, 2018)

Verbal Ellipsis

هو يعشق الشتاء وأنت ~~تعشق~~ الخريف. (ArTenTen)

He adores the winter, and you ~~adore~~ the fall?

ويجز بيضحي بكاريره ومصر كمان ~~بتضحي~~ (Twitter)

Wegz is sacrificing his career; and Egypt is also ~~sacrificing.~~

A verb phrase is omitted or elided from a syntactic construction, contingent upon the presence of its antecedent within the immediate linguistic context. In other words, VP ellipsis always targets an entire VP, which usually occurs where two clauses are coordinated, and an equivalent VP exists in the other clause.

(Carnie, 2021; Cannon, 2023)

Gapping

أنت دخلت من الباب ونحن ~~دخلنا~~ من النافذة (ArTenTen)

You entered from the door, and we ~~entered~~ from the window.

هو دخل الكلية الحربية وانا ~~دخلت~~ هندسة (Twitter)

He joined the military school, and I ~~joined~~ engineering.

For **Gapping** to happen in Arabic, we need three conditions:

1) There must be surrounding lexical material on both sides of the elided verb in the second conjunct.
2) Constituents, after the verb in the second conjunct and in the first conjunct, must be syntactically and semantically parallel.
3) At least two remnants must be left behind.

(Mansour, 2007)

Stripping

تركز عليها أغلب الدول والشركات كذلك ~~تركز عليها~~ (ArTenTen).

Most countries focus on it, and companies also ~~focus on it~~.

أبو تريكة اعتزل في عزه وبركات كمان ~~اعتزل في عزه~~ (Twitter)

Abo Treka retired in his prime, and Barakat also ~~retired in his prime~~.

In **Stripping**, an entire clause is omitted except for one constituent, i.e., the remnant. In some Arabic dialects, e.g., Libyan and Iraqi, it is argued that for stripping to happen there has to be a sentential modal adverb such as probably or maybe along with a focusing adverb such as 'too', with the latter being the only obligatory condition.

(A. Algryani, 2013; Albuarabi, 2019)

Sluicing

نعم السودان يمكنه أن يحقق طفرات، ولكن كيف ~~يحقق طفرات~~؟ (ArTenTen).

Yes! Sudan can make some breakthroughs but how ~~can Sudan make some breakthroughs~~?

مفيش مشكلة نستحمل بس ليه ~~نستحمل~~؟ ولامتى ~~نستحمل~~؟(Twitter)

OK! We can endure this, but why ~~endure it~~?

And for how long ~~should we endure~~?

Sluicing represents a type of surface anaphora that requires an antecedent, i.e., the omitted content aligns with the content of a sentence within the discourse. Also, the wh-phrase that has been sluiced is understood as a completely pronounced wh-question, as it carries the complete interrogative force and is consequently equivalent to fully expressed wh-question.

(Merchant, 2006; A. Algryani, 2019)

Fragment Answer

ما الذي يجب أن يكون جزءا من منهج طفلك؟ العلوم ~~يجب أن تكون جزءا من منهج طفلي~~.

(ArTenTen)

What should be in your child's curriculum? Science ~~should be in my child's curriculum~~.

تحب تسافر فين؟ لندن! (Twitter) ~~أحب اسافر~~

Where would you like to travel? ~~I would like to travel to~~ London.

Fragment answers are short answers to questions consisting of non-sentential XPs. These XPs, although lacking a full sentence structure, convey the same propositional content as full sentential answers.

(A. Algryani, 2017)

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%**
    - Precision 0.56, Recall 0.18, Accuracy 72%
  - Task 3: 0-shot word filling
    - Accuracy ~80%

Experiments

English in comparison:

- Task 1:
  - Logistic Regression (baseline): accuracy 72%
  - 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~~!