Natural Language Generation in the Logos Model

Sara Amato & Kutz Arrieta

2023

NATURAL LANGUAGE GENERATION	GENERATIVE AI
can be defined as the means and methods to produce human language from another language, from coded instructions, from graphical representations, from data collections	is a type of artificial intelligence technology that can produce various types of content including text, imagery and audio exactly like Natural Language Generation, what differs is the methodology.

Natural Language Generation

Applications

To create language models that are easily understood by humans not requiring them to have specific knowledge of a certain domain. (*Model-Driven Engineering - Håkan Burden and Rogardt Heldal*)

To generate reports from data sets (*Arria*)

To produce both questions and explanations from Natural Language understanding and reasoning systems. (*KnowMatters or IBM Watson*)

All based on different approaches to natural language

Natural Language Generation

Models

Statistical models

MT systems based solely on statistical or neural models, with no, or very limited, semantic generation

Match patterns in aligned bilingual texts to build a statistical model of translation

Not comparable to the tasks of parsing and generation in systems based on linguistic knowledge

Natural Language Generation

Models

Dependency Grammars

Sentence generation is viewed as a sequence of transductions (surface representations), produced by different grammars

The Universal Networking Language

The Universal Networking Language (UNL) has been used as a tool in generation systems.

Logos uses its own SAL language for all its modules.

Natural Language Generation

Projects

Wikimedia Abstract Language project

The main goal is to make it possible for less-resourced languages to generate content through community users.

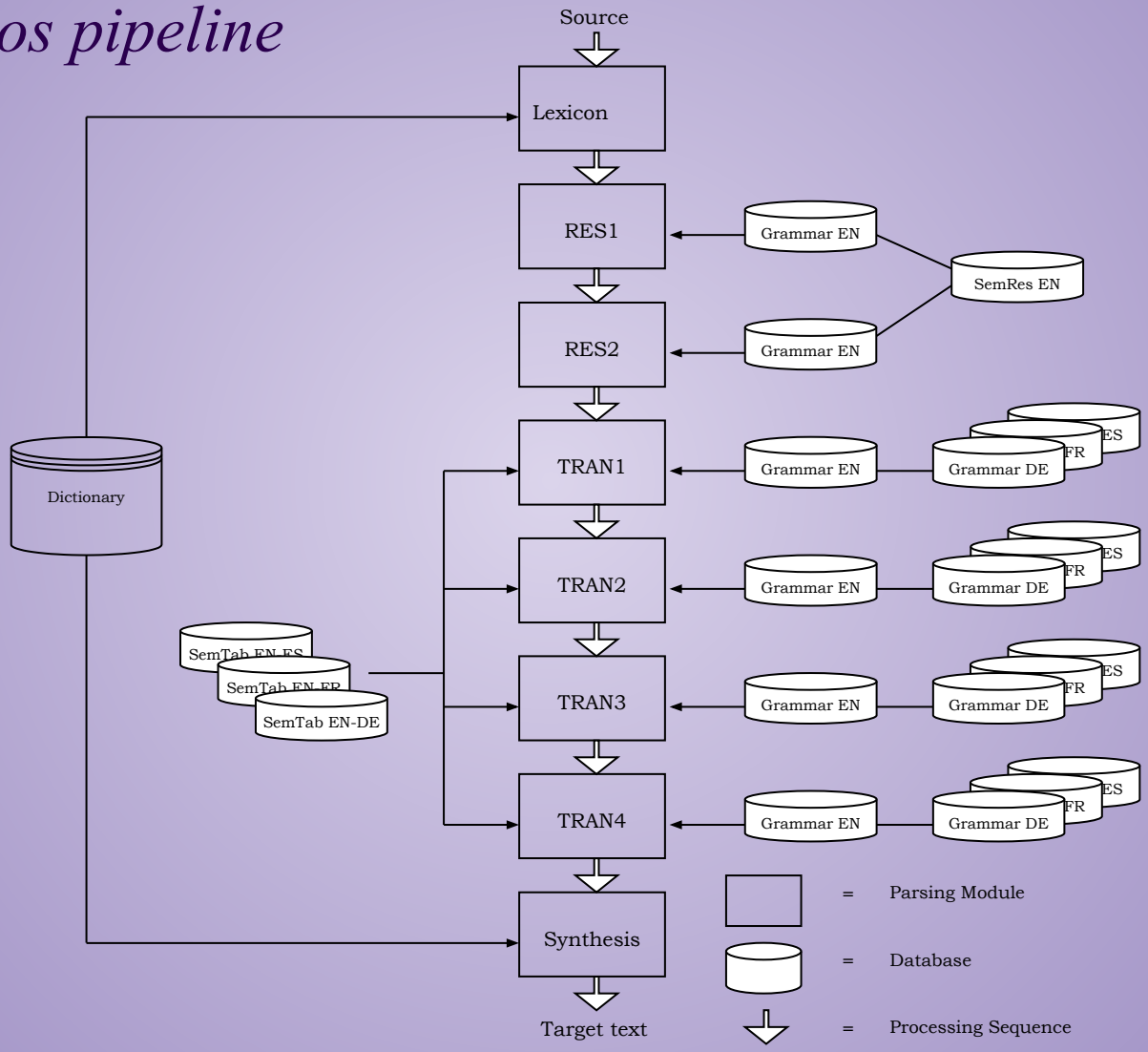
LLMs are not being adopted:

those languages do not have enough digital content for the models to be trained on

Maria Keet and her team in University of Cape Town on isiZulu languages.

The Logos Model

The Logos pipeline



Translation System Architecture

RES

Resolves homograph ambiguity

Segments the source sentence into clauses

Translation System Architecture

Tran1 - Tran4

Four modules work together to produce a bottom-up parse

Call target language-specific components which perform transfer operations

Translation System Architecture

Tran1-Tran4

Tran1

Analyzes Simple Noun Phrases (e.g., Article + Noun) and Verb Phrases (e.g., Auxiliary + Verb)

Tran2

Analyzes more Complex Noun Phrases (e.g., Noun + Prep + Noun, Noun + Relative Clauses)

Tran3

Analyzes Verb Argument Structure and Determines Final Prepositional Phrase Attachment

Tran4

Determines Relationships between Clauses and Establishes Target Sentence Word Order

The Logos Model

TRAN1 already has rewrite rules (that transform source language structures into target language structures).

Source analysis rules trigger generation rules.

TRAN4 is where authentic generation rules occur and well-formed sentences in the target language are built.

TRAN4, the generation module, is supposed to be multi-source and not dependent on a particular source analysis of one specific language or another. It is based on an abstract representation or interlingua.

The Logos Model

SAL

SAL is the semantico-syntactic representation to encode languages and rules

Based on a higher level of abstraction which goes beyond the usual part-of-speech classifications.

Reflects the deep semantic functionality of each part of speech, whereby different members of a word class belong to a similar semantic category provided that they trigger similar syntactic behavior.

The Logos Model

SAL

send and **give** have identical chains of semantico-syntactic codes

a. Their deep semantics calls for a second indirect object
b. The indirect object can be introduced by the preposition ***to*** when following the direct object (***he gave a camera to his wife***) or by no preposition at all by inverting the word order of the two objects (***he gave his wife a camera***).

communicate, instead, shares only part of its semantico-syntactic chain with ***send*** and ***give***

Its syntactic environment only shares with them a) but not b).

The Logos Model

Morphological modules

Address both parsing and generation

Need to encode all the information necessary to function both in the context of a source and a target language.

The rules need to be much more precise to avoid spurious tokens.

The Logos Model

Transfer parse and generation

The source tree is built and rebuilt to accommodate the needs of the target language.

TRAN rules are target or group-specific, and call target-specific tables (30-tables, 40-tables and 50-tables), which accomplish different tasks, getting more and more specific to a given target.

Then the generation phase takes place, where constituent movement, lexical selection and final formatting take place.

The Logos Model

Evolution

The Logos model has been evolving since its conception towards a greater separation of source and target modules aiming at a full Interlingua model

This change in design is motivated by the need for modularity in order to improve results and accelerate the addition of new language pairs.

The Logos Model

Evolution

Greater separation of source and target modules aiming at a full Interlingua model

Improving results and accelerating the addition of new language pairs

Completing source operations independently of the target language

Target languages concentrating on generation from a SAL parse tree, without any concern of impact on the source language parse or on other target languages

Linguistic Challenges in Generation

Generation of Verb Phrases and Verb Compounds

SEMTAB rules handling "verb + particle" structures

LOOK (VI) OUT (PART) = TENER CUIDADO

LET (VT) OFF (PART) = DESPEDIR

RES module has resolved that an element is a particle and not a preposition or an adverb therefore the combination of the "verb + particle" strings in a rule represents a different verb with a different semantico-syntactic code from the one assigned to the original main verb, and a different transfer in the target language.

Linguistic Challenges in Generation

Generation of Verb Phrases and Verb Compounds

In German source, separable verb prefixes and particles must be reassigned to the verb so that they can be handled as a single string.

Wir drehen weiter

each word enters the translation module separately, **drehen** and **weiter** would be handled separately

once RES confirms that weiter should be treated as a separable particle there will be a match on the rule
DREHEN WEITER = CONTINUARE A GIRARE

Linguistic Challenges in Generation

Semantic context

SEMTAB rules handling “verb + noun” patterns

BE (VI) (UNITS OF LINEAR MSR-PREC BY ARITHMATE) = MEDIR N

These rules handle as an idiom the combination of the copulative verb and a set of any noun that belong to a certain semantic category, under any form or any modification context, and tailoring the translation appropriately.

Linguistic Challenges in Generation

Quantifiers

EN - any two books ***ES***
- dos libros cualesquiera

The default transfer for this phrase would have been ****cualquier dos libros***, but a TRAN rule, dealing with the source noun phrase analysis and sending a signal to the transfer module causes the Generation module to effect the correct output.

Linguistic Challenges in Generation

Clitics

EN - You may contact him

The Logos model produces ***se puede poner en comunicación con él***

Source analysis takes care of choosing between ***le*** and ***con él,*** etc. in different contexts. In this specific example, a SEMTAB rule is making the decision:

CONTACT (VT-ACTIVE) N(NOM-HUMAN) N = N
PONERSE(REFL) EN COMUNICACIÓN CON N

Linguistic Challenges in Generation

The black hole strategy

Decides whether an element goes just after the head of a phrase.

You you may ask him → gli potete chiedere

you can always give it to your teacher → lo potete sempre dare al vostro insegnante

Black holes apply to both Spanish and Italian and can also be used in all types of phrases.

Linguistic Challenges in Generation

Additional examples where the Generation module can make the correct decision when source parsing carries enough information:

-ed in English

ser and estar /essere and stare

existentials

ellipsis

adjective Ordering

elision in Italian

determiners

The Logos Model

Conclusion and Future Work

Target Independent Analysis (TIA) is the logical next step in the Logos Model.

Modularity and independent linguistic work

Intermediate system where additional source analysis operations might be performed for the sake of the Generation module

Target SEMTAB and target verb valence information could be encoded, providing the Generation module with very powerful tools

The Logos Model

Conclusion and Future Work

The Logos Model should also find a way to integrate statistical and neural models.

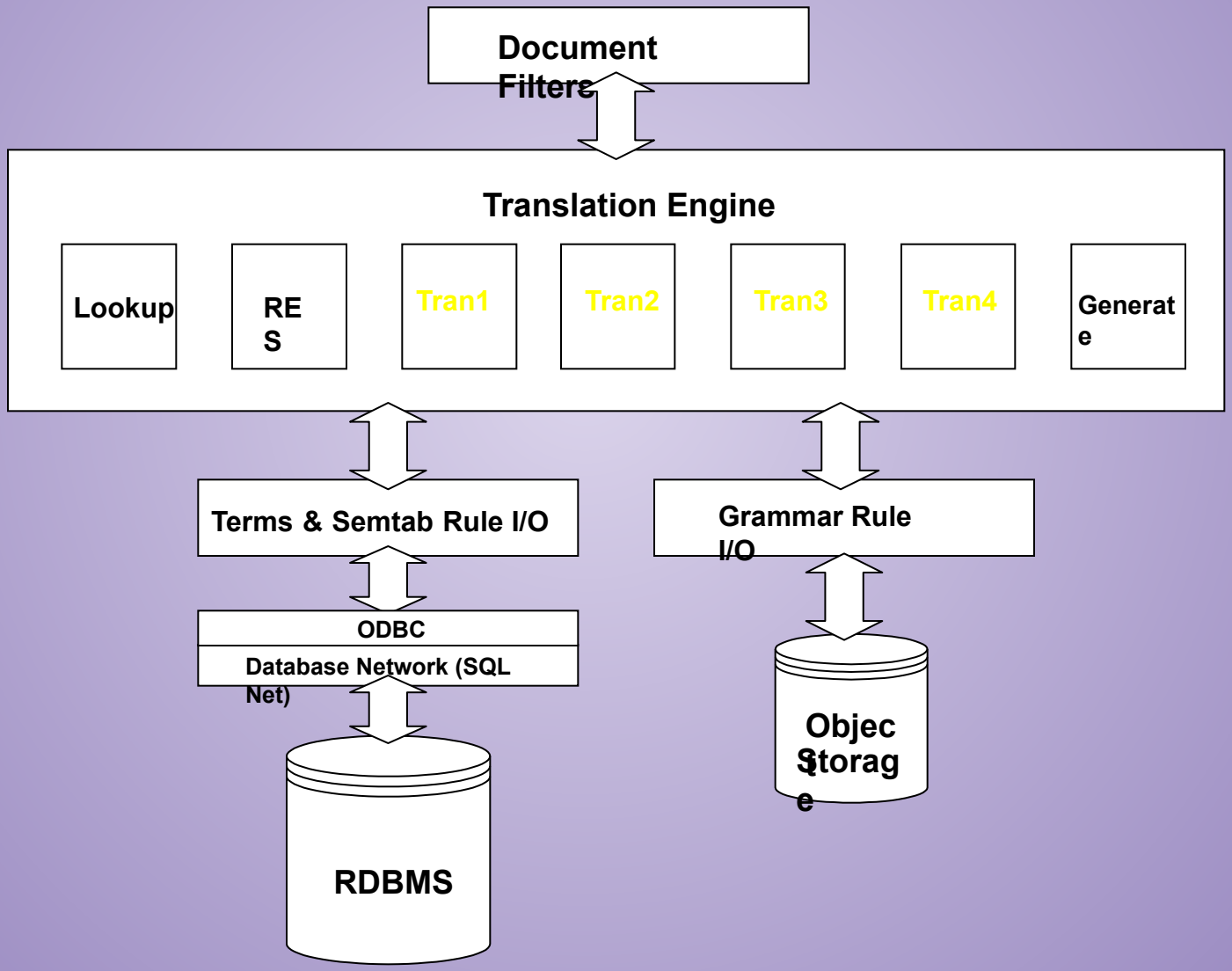
Combining the power of both strategies could make the Logos Model the best performing system in the market

Thank you!

Questions?

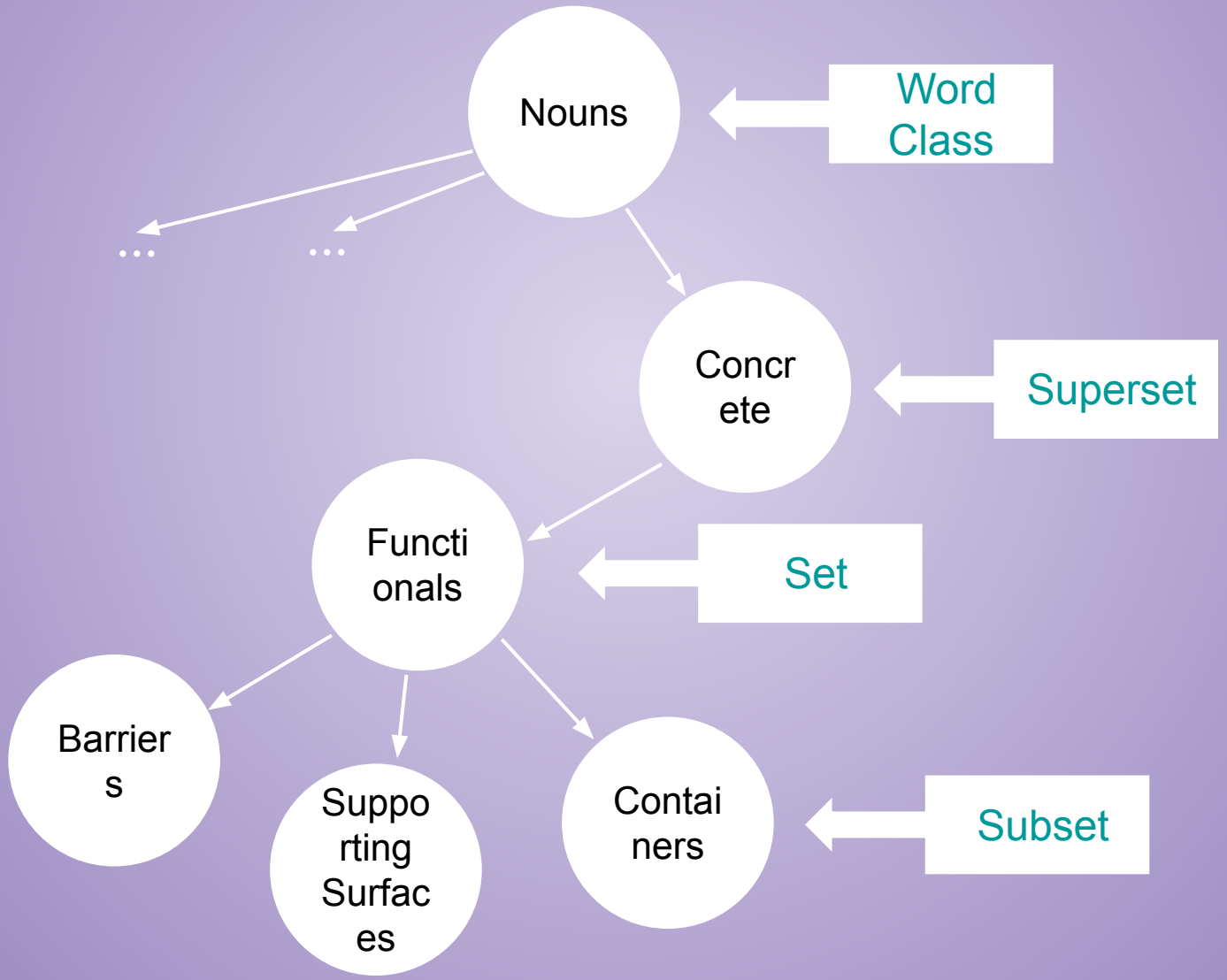
The Logos Model

Translation System Architecture



The Logos Model

SAL Ontology for Nouns



The Logos Model

Motivation for SAL

- **The Power of Analogy**
  - Example:

    bin table

    pocket = container shelf = supporting surface

    drawer floor

    box bench

    from table, chair, bench → *off of* N

    from **box, drawer, pocket** → *out of* N

The Logos Model

Motivation for SAL

Syntactic Disambiguation using SAL

Example:

ways of cooking lentils (ING=V)

types of cooking utensils (ING=AJ)

SAL Codes of the first noun determines analysis of ING form

The Logos Model

Motivation for SAL

Semantic Disambiguation using SAL

Example:

run a machine → *far funzionare*

run a company → *dirigere*

run a program → *eseguire*

Disambiguation is accomplished by semantic rules

The Logos Model

Motivation for SAL

Monotonic SAL Environment:

- Source language sentence and linguistic patterns in rule bases expressed as SAL elements
- Source sentence SAL elements serve as search arguments to the rule bases

SAL rules interact on the source sentence elements

- Determine the structure and meaning of the source language sentence
- Construct an equivalent sentence in the target language