Making Choices
Statistical Microplanning

Claire Gardent

CNRS/LORIA and Université de Lorraine, Nancy

XRCE, October 2015
Grenoble, France

Joint Work with

Laura Perez-Beltrachini

Funded by the French ANR Project WebNLG
http://talc1.loria.fr/webnlg/stories/about.html

Writing/Producing a Text = Making Choices

What is talked about ? (Content Selection)

Structuring the selected data into a text plan (Document planning)

Producing fluent text (Microplanning)

- Describing entities (Generating Referring Expressions)
- Choosing lexical items and syntactic structures (Lexicalisation, Surface Realisation, Aggregation, Sentence Segmentation)

Outline

1. Generating from Knowledge-Bases
2. NLG Approaches
3. A Grammar-Based Statistical Approach for Microplanning

Semantic Web and Knowledge-Bases

Ontologies

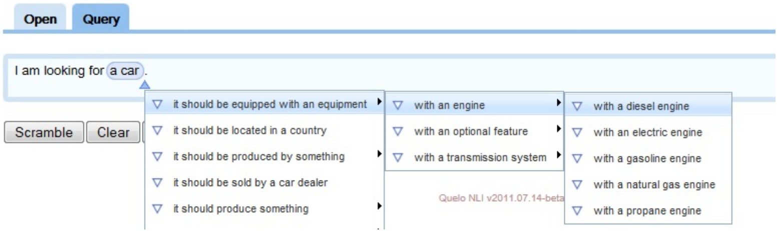
- Biomedical domain: SNOMED, GO, BioPAX, the Foundational Model of Anatomy and the U.S. National Cancer Institute Thesaurus
- Ontologies for e.g., geography, geology, agriculture and defence

Large scale RDF datasets

- DBPedia, Geonames, US Census, EuroStat, MusicBrainz, BBC Programmes, Flickr, DBLP, PubMed, UniProt, FOAF, SIOC, OpenCyc, UMBEL, Yagoo ...

Generating from Knowledge-Bases

Many applications could benefit from KB-to-Text generation.



Quelo Natural Language Interface

(a) Natural Language Interfaces for KB

Generating from Knowledge-Bases

Many applications could benefit from KB-to-Text generation.

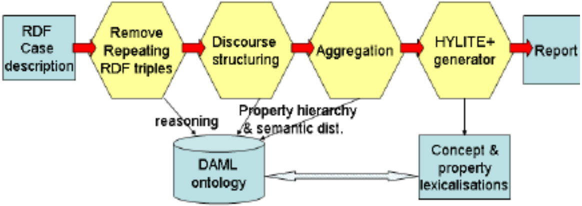


Fig. 1. The MIAKT Generator

The Miakt System: Generating Patient Report from RDF data

(b) Natural Language Descriptions of KB entities/concepts

Generating from Knowledge-Bases

Many applications could benefit from KB-to-Text generation.

Class label	OWL axioms (Manchester syntax)	Natural Language Definition Extracted
22rv1	bearer_of some 'prostate carcinoma' derives_from some 'Homo sapiens' derives_from some prostate	A 22rv1 is a cell line. A 22rv1 is all of the following: something that is bearer of a prostate carcinoma, something that derives from a homo sapiens, and something that derives from a prostate.
HeLa	bearer_of some 'cervical carcinoma' derives_from some 'Homo sapiens' derives_from some cervix derives_from some 'epithelial cell'	A he la is a cell line. A he la is all of the following: something that is bearer of a cervical carcinoma, something that derives from a homo sapiens, something that derives from an epithelial cell, and something that derives from a cervix.
Ara-C-resistant murine leukemia	has subclass b117h* has subclass b140h*	A ara c resistant murine leukemia is a cell line. A b117h, and a b140h are kinds of ara c resistant murine leukemias.
GM18507	derives_from some 'Homo sapiens' derives_from some lymphoblast has_quality some male	A gm18507 is all of the following: something that has as quality a male, something that derives from a homo sapiens, and something that derives from a lymphoblast.

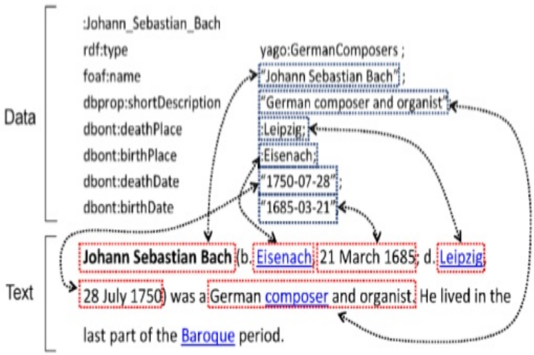
The SWAT System: Verbalising KB Content

(c) Natural Language Presentation of KBs

Natural Language Generation

Manually or Automatically Acquired Templates (Duma et al. 2010, Blake et al. 2013, Schilder et al.2013)

Align Text and Data



Create Template

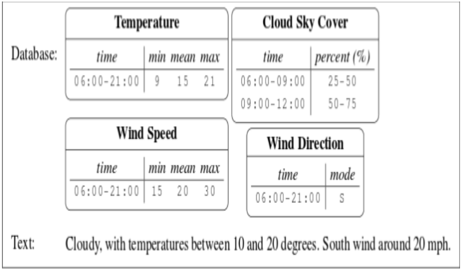
Name (b.birthPlace, birthDate, d. deathPlace, deathDate) was a shortDescription.

Picture from Duma and Klein 2010

Natural Language Generation

Use Machine Learning to map KB Data to NL Phrases (Wong et al. 2007, Belz 2008, Angeli et al. 2010, Chen et al. 2008, Konstas and Lapata 2012a , Konstas and Lapata 2012b)

Parallel Corpus



Database:

Temperature

time	min	mean	max
06:00-21:00	9	15	21

Cloud Sky Cover

time	percent (%)
06:00-09:00	25-50
09:00-12:00	50-75

Wind Speed

time	min	mean	max
06:00-21:00	15	20	30

Wind Direction

time	mode
06:00-21:00	S

Text: Cloudy, with temperatures between 10 and 20 degrees. South wind around 20 mph.

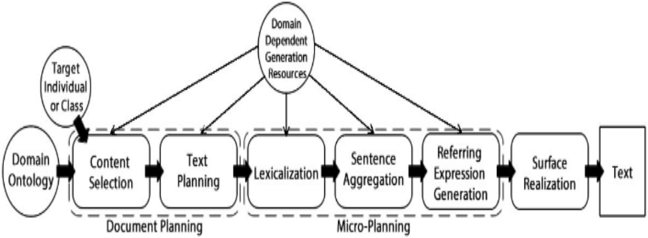
Picture from Konstas and Lapata 2013

Learn Mapping

- Probabilistic CFG mapping DB to Text
- Cascaded Discriminative models
- Statistical Machine Translation

Natural Language Generation

Use hand-crafted lexicon, grammar and text plans (Dimitrios et al. 2007, Androtsopoulos et al. 2013, Power et al 2010, Bontcheva et al 2004)



The NaturalOwl System: Describing individuals or classes of owl ontologies

KB to Text Generation

Parallel Data-to-Text corpus is hard to get

KB to Text Generation

Parallel Data-to-Text corpus is hard to get

Manually crafted grammars, lexicons and text plans are costly to develop

KB to Text Generation

Parallel Data-to-Text corpus is hard to get

Manually crafted grammars, lexicons and text plans are costly to develop

Microplanning problem: grammatical $\neq$ fluent

KB to Text Generation

Parallel Data-to-Text corpus is hard to get

Manually crafted grammars, lexicons and text plans are costly to develop

Microplanning problem: grammatical $\neq$ fluent

? I am looking for a flight. Its departure date should be November 5th. The arrival date of the flight should be November 6th. The destination of the flight should be Paris.

✓I am looking for a flight whose departure date should be November 5th, whose arrival date should be November 6th and whose destination should be Paris.

A Statistical Grammar-Based Approach

Input = KB Query

A Statistical Grammar-Based Approach

Input = KB Query
Segment Input, lexicalise KB symbols, aggregate and realise

A Statistical Grammar-Based Approach

Input = KB Query
Segment Input, lexicalise KB symbols, aggregate and realise

$\texttt{Professor} \sqcap \texttt{Researcher} \sqcap \exists\texttt{teach.LogicCourse}$
$\sqcap \exists\texttt{worksAt.AlicanteUniversity}$

A Statistical Grammar-Based Approach

Input = KB Query
Segment Input, lexicalise KB symbols, aggregate and realise

$\texttt{Professor} \sqcap \texttt{Researcher} \sqcap \exists\texttt{teach.LogicCourse}$
$\sqcap \exists\texttt{worksAt.AlicanteUniversity}$

I am looking for a professor who is a researcher and teaches a course on logic. He should work for Alicante University.

A Statistical Grammar-Based Approach

Combines a grammar, a lexicon with a surface realisation algorithm integrating a hypertagger, a beam search and a ranker

A Statistical Grammar-Based Approach

Combines a grammar, a lexicon with a surface realisation algorithm integrating a hypertagger, a beam search and a ranker

The grammar

- Defines the space of possible realisations
- Enforces hard constraints (grammaticality)

The Statistical Modules (Hypertagger, Beam Search, Ranker)

- Allow for efficiency (speed)
- Enforce soft constraints (fluency)

The Generation Algorithm

- Hypertagging: Selects the n-best sequences of grammar rules (TAG trees) given the input semantics
- Lexical Selection: retrieves TAG trees whose semantic subsumes the input and which are compatible with the hypertagger decisions
- Surface Realisation: Combines TAG trees to produce Sentences
- Ranking: Select n best outputs using Language Model

Grammar Based Generation

Input = KB Query

$\texttt{Professor} \sqcap \texttt{Researcher} \sqcap \exists\texttt{teach.LogicCourse}$
$\sqcap \exists\texttt{worksAt.AlicanteUniversity}$

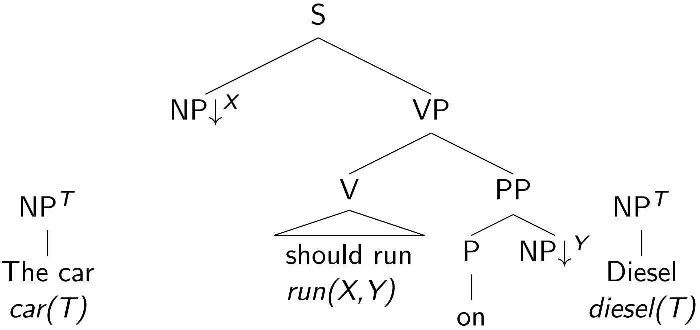
Grammar Based Generation

Input = KB Query

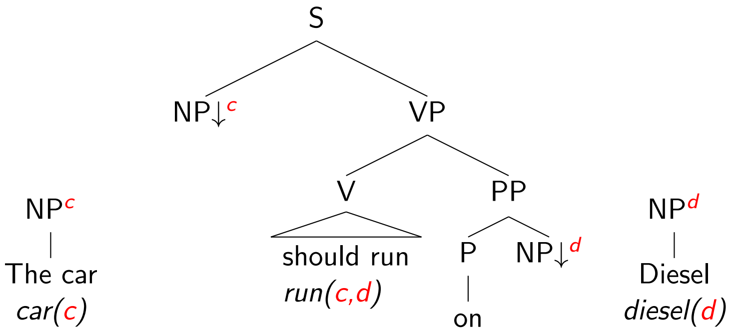
$\texttt{Professor} \sqcap \texttt{Researcher} \sqcap \exists\texttt{teach.LogicCourse}$
$\sqcap \exists\texttt{worksAt.AlicanteUniversity}$

`Professor(p) Researcher(p) teach(p c) LogicCourse(c) worksAt(p u) AlicanteUniversity(u)`

Grammar Based Generation

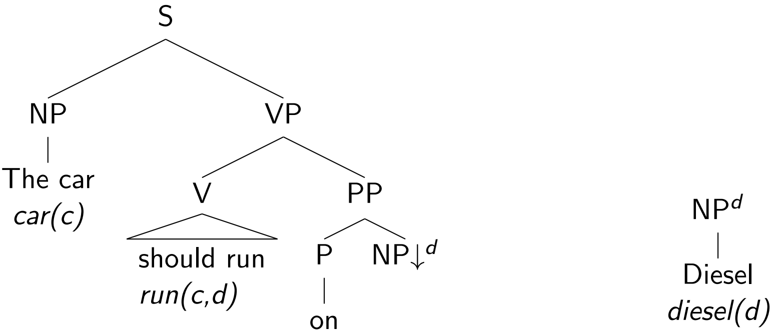


Grammar Based Generation



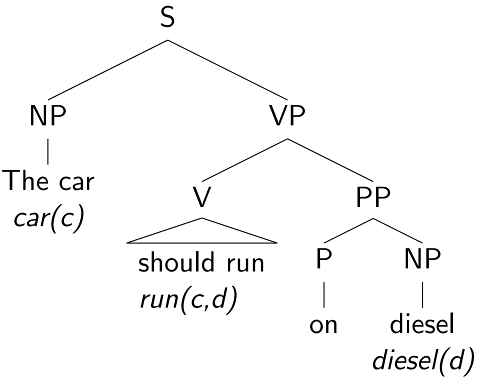
car(c), run(c,d), diesel(d)

Grammar Based Generation



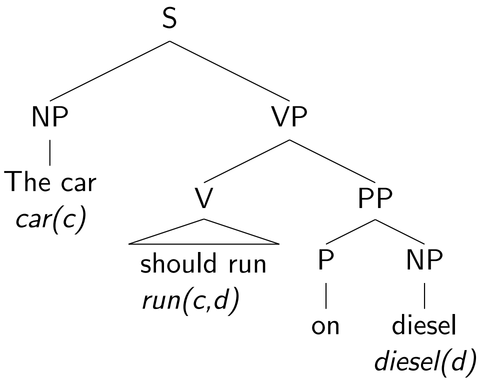
car(c), run(c,d), diesel(d)

Grammar Based Generation



car(c), run(c,d), diesel(d)

Grammar Based Generation



`The car should run on diesel`

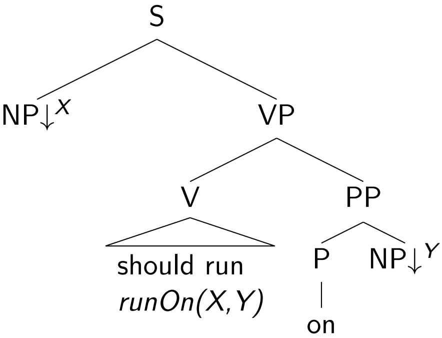
Grammar-Based Generation

The lexicalised grammar is very big (n * number of words) (tractability)

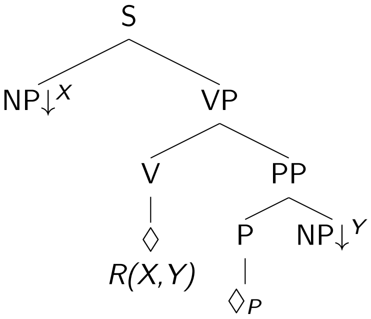
Grammar-Based Generation

The lexicalised grammar is very big (n * number of words) (tractability)
→ Separate Grammar from Lexicon

Grammar and Lexicon

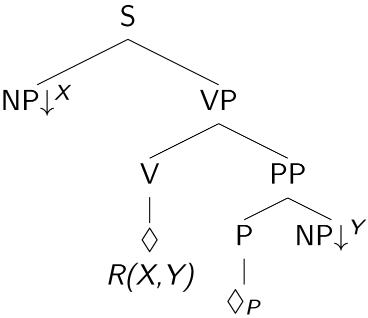


Grammar and Lexicon



Semantics	*runOn*
Tree	nx0Vpnx1
Anchor	should run
Co-Anchor	P $\rightarrow$ on

Grammar and Lexicon



Semantics	*runOn*
Tree	nx0Vpnx1
Anchor	should run
Co-Anchor	P $\rightarrow$ on

Semantics	*assistWith*
Tree	nx0Vpnx1
Anchor	should assist
Co-Anchor	P $\rightarrow$ with

Grammar and Lexicon

The lexicon

- relates KB Symbols, Natural Language Expressions and Syntax (Grammar rules). It is domain specific.
- is acquired automatically

Grammar and Lexicon

The lexicon

- relates KB Symbols, Natural Language Expressions and Syntax (Grammar rules). It is domain specific.
- is acquired automatically

The grammar

- specifies the various syntactic realisations of words. It is generic.
- is a small, manually specified Tree Adjoining Grammar

Automatic Lexicon Induction

The lexicon is automatically derived from KB symbols (Trevisan 2010)

Automatic Lexicon Induction

The lexicon is automatically derived from KB symbols (Trevisan 2010)

Step 1: Tokenize and PoS Tag

`runsOn` $\rightarrow$ runs/VBD on/IN

Automatic Lexicon Induction

The lexicon is automatically derived from KB symbols (Trevisan 2010)

Step 1: Tokenize and PoS Tag

`runsOn` $\rightarrow$ runs/VBD on/IN

Step 2: The result sequence is mapped to one or more Lexical Entries

Automatic Lexicon Induction

The lexicon is automatically derived from KB symbols (Trevisan 2010)

Step 1: Tokenize and PoS Tag

`runsOn` $\rightarrow$ runs/VBD on/IN

Step 2: The result sequence is mapped to one or more Lexical Entries

runs/VBD on/IN $\rightarrow$

Semantics	*runsOn*
Tree	nx0Vpnx1
Anchor	should run
Co-Anchor	P $\rightarrow$ on

Generic Grammar

A small (100 trees), hand-written generic grammar models subcategorisation and syntactic variation.

Syntactic Variations

NP_0 should be equipped with NP_1	Canonical
and NP_0 should be equipped with NP_1	S-Coordination
NP_0 which should be equipped with NP_1	SubjRel
NP_0 (...) and which should be equipped with NP_1	SubjRelPU
NP_0 (...), which should be equipped with NP_1	SubjRelPU
NP_0 equipped with NP_1	PpartOrGerund
NP_0 (...) and equipped with NP_1	SharedSubj
NP_0 (...), equipped with NP_1	SharedSubj
NP_1 with which NP_0 should be equipped	PObjRel
NP_0 (equipped with X) and with NP_1	Ellipsis
NP_0 (equipped with X), with NP_1	Ellipsis

Generic Grammar

A small (100 trees), hand-written generic grammar models subcategorisation and syntactic variation.

Valency/Subcategorisation Variations

NP_0 should generate NP_1	nx0VVnx1	Canonical
NP_0 should run on NP_1	nx0VVpnx1	Canonical
NP_0 should be equipped with NP_1	nx0VVVpnx1	Canonical
NP_0 should be the equipment of NP_1	nx0VVDNpnx1	Canonical
NP_0 should have access to NP_1	nx0VVNpnx1	Canonical
NP_0 should be relevant to NP_1	nx0VVApnx1	Canonical
NP_0 should be an N_1 product	nx0VVDNnx1	Canonical
NP_0 with NP_1	betanx0Pnx1	Canonical

Making Choices (Hypertagging)

The hypertagger

- Filters the initial search space (efficiency)
- Is trained to eliminate sequences of grammar trees that lead to less fluent sentences (fluency)

Making Choices (Hypertagging)

Output of the Lexical Selection

CarDealer(X)	locatedIn(X,Y)	City(Y)	sell(Y,Z)	Car(Z)	runOn(Z,W)	Diesel
nx	nx0VVVpnx1	nx	nx0VVVnx1	nx	nx0VVpnx1	nx
	PRO0VVVpnx1		PRO0VVVnx1		...	
	sCONJnx0VVVpnx1		sCONJnx0VVVnx1		...	
	sCONJPRO0VVVpnx1		sCONJPRO0VVVnx1		...	
	W0nx0VVVpnx1		W0nx0VVVnx1		...	
	ANDWHnx0VVVpnx1		ANDWHnx0VVVnx1		...	
	COMMAWHnx0VVVpnx1		COMMAWHnx0VVVnx1		...	
	betanx0VPpnx1		betanx0VPpnx1		...	
	betanx0ANDVPpnx1		betanx0ANDVPpnx1		...	
	betanx0COMMAVPpnx1		betanx0COMMAVPpnx1		...	
	W1pnx1nx0VV		W1pnx1nx0VV		...	
	betavx0ANDVVVpnx1		betavx0ANDVVVnx1		...	
	betavx0COMMAVVVpnx1		betavx0COMMAVVVnx1		...	

I am looking for a car dealer located in a city who should sell cars. The car should run on diesel.

Making Choices (Inversed Parsing and Ranking)

The **hypertagger** prunes the initial search space and favours Tree/Syntactic Classes sequences which yield fluent sentences.

$\texttt{CarDealer} \sqcap \exists \texttt{locatedIn.City} \sqcap \exists \texttt{sell.Car} \sqcap \exists \texttt{runOn.Diesel}$

Making Choices (Inversed Parsing and Ranking)

The **hypertagger** prunes the initial search space and favours Tree/Syntactic Classes sequences which yield fluent sentences.

`CarDealer` $\sqcap$ $\exists$`locatedIn.City` $\sqcap$ $\exists$`sell.Car` $\sqcap$ $\exists$`runOn.Diesel`

Tbetanx0VPpnx1 TANDWHnx0VVnx1 Tnx0VVpnx1 Tnx
I am looking for a car dealer located in a city and who should sell a car. The car should run on diesel.

~~Tnx0VPpnx1 Tnx0VVnx1 Tnx0VVpnx1~~
~~*I am looking for a car dealer. He should be located in a city. He should sell a car. The car should run on diesel.*~~

Making Choices (Hypertagging)

Hypertagging

A linear-chain Conditional Random Field model is used to define the posterior probability of labels (TAG trees, syntactic classes) $y=\{y_1, \ldots, y_n\}$ given features informed by the input semantics $x=\{x_1, \ldots, x_k\}$:

$$P(y \mid x) = \frac{exp[\sum_{j=1}^{m} \sum_{i=1}^{n} \theta_j f_j(x, i, l_i, l_{i-1})]}{\sum_{l'} exp[\sum_{j=1}^{m} \sum_{i=1}^{n} \theta_j f_j(x, i, l'_i, l'_{i-1})]}$$

Given a set of candidate hypertags (TAG trees) associated with each literal, the hypertagger finds the optimal hypertag sequence y^* for a given input semantics x:

$$y^* = argmax_y P(y \mid x)$$

Experimental Setup

Grammar and Lexicon

- Grammar: 69 trees, 10 syntactic classes
- Lexicon: 13 KB, 10K entries, 1296 concepts and elations, average lexical ambiguity: 7.73.

Evaluation Metrics

- Hypertagging Accuracy
- Coverage and Speed
- Output quality (Human Evaluation)
- Qualitative Analysis (Microplanning)

Comparison Models

- Template-Based Model
- Symbolic Grammar-Based Model

Data

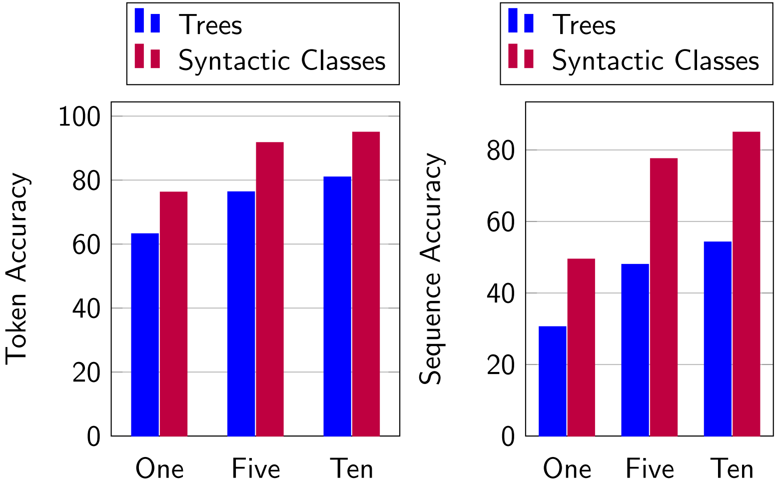
Training Data for the CRF

- 206 training instances = (KB query, tree sequence) pairs
- From 11 ontologies (Domain Independent)
- Input Length (min:2, max:19, avg: 7.44)
- CRF trained and tested using 10 fold cross validation

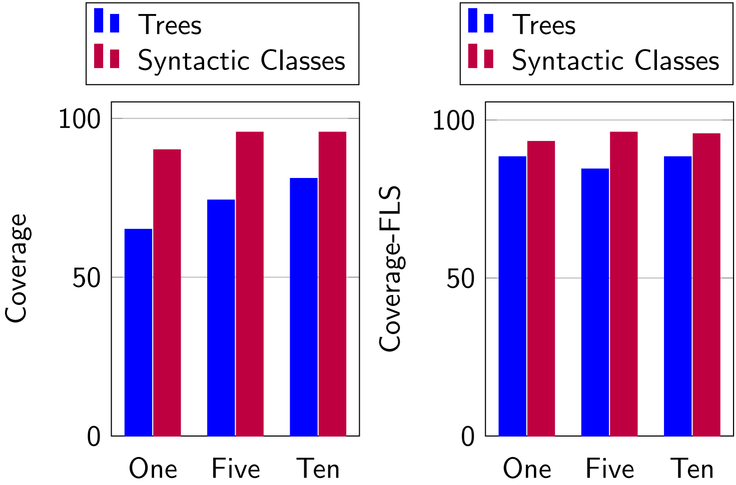
Features

- KB Symbol: Shape and content (words) of relation names (unigram and bigrams)
- Lexical features: word overlap between KB symbols, presence/absence of prepositions, etc.
- Entity Chaining Features: distribution of discourse entities in the input query
- Structural features: length of the input, number of predications over the same entity ...

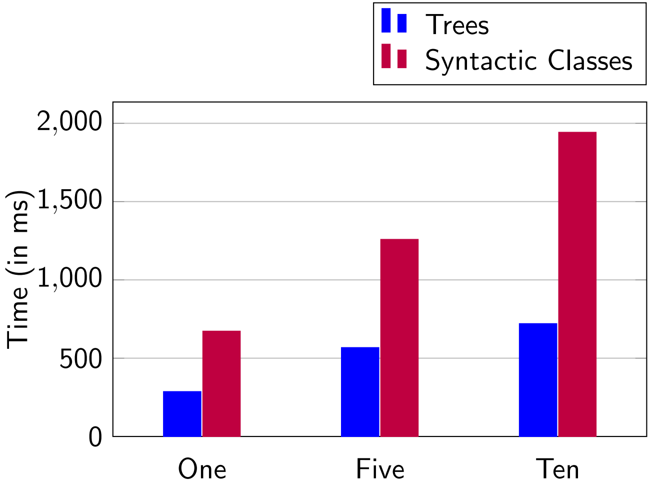
Results: Hypertagging Accuracy



Results: Coverage



Results: Speed

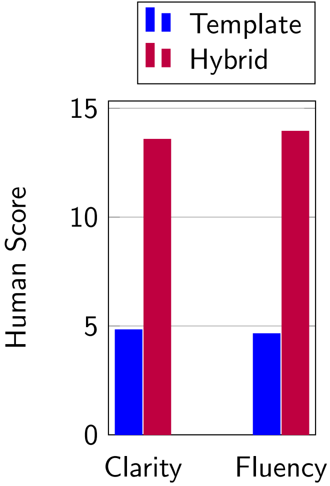


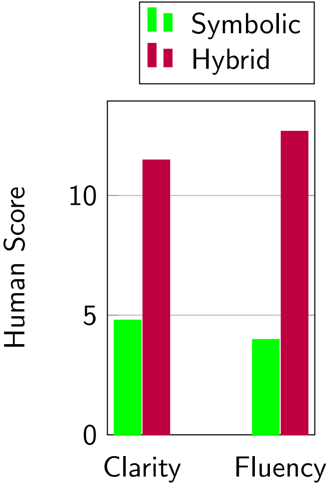
Results: Output quality

Human Evaluation

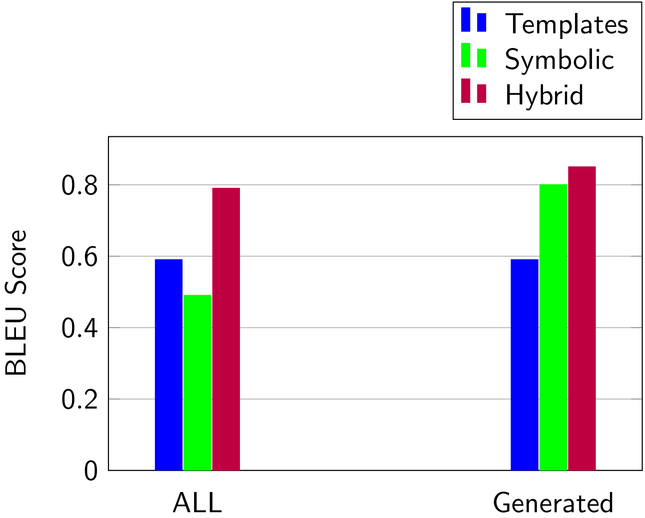
- 48 input queries
- from 13 knowledge bases (2 not used in training corpus)
- 24 raters
- Online evaluation
- Sliding ruler
- Scale 0-50
- Latin Square design

Results: Output quality





Results: Output quality (BLEU Scores)



Example Ouput: Sentence Segmentation

3 relations, 4 concepts: 1 sentence

I am looking for a used car whose color should be white, which should be located in a France and whose model should be a toyota 4 runner.

Example Ouput: Sentence Segmentation

3 relations, 4 concepts: 1 sentence

I am looking for a used car whose color should be white, which should be located in a France and whose model should be a toyota 4 runner.

4 relations, 5 concepts: 2 sentences

I am looking for a new car whose exterior color should be beige and whose body style should be a utility vehicle. The new car should run on a natural gas and should be located in a country.

Example Ouput: Sentence Segmentation

3 relations, 4 concepts: 1 sentence

I am looking for a used car whose color should be white, which should be located in a France and whose model should be a toyota 4 runner.

4 relations, 5 concepts: 2 sentences

I am looking for a new car whose exterior color should be beige and whose body style should be a utility vehicle. The new car should run on a natural gas and should be located in a country.

3 relations, 5 concepts: 2 sentences

I am looking for a new car whose body style should be a utility vehicle, an off road. The new car should run on a natural gas and should be located in a country.

Example Output: Syntactic Variation

I am looking for a car dealer **located in a country** *and who should sell a car whose make should be a toyota. The car should run on a fuel and should be equipped with a manual gear transmission system.* *(Participial)*

I am looking for a car dealer who should sell a new car whose model should be a toyota. **It should be located in a country**. *(VP with pronominal subject)*

I am looking for a new car, an off road whose body style should be a utility vehicle. The new car should run on a natural gas **and should be located** *in a country.* *(Coordinated VP)*

I am looking for a car produced by a car make. The car make should be the make of a toyota. The car make **should be located** *in a city and should produce a land rover freelander.* *(Canonical Declarative Sentence)*

Example Output: Aggregation

VP Coordination

NewCar (...) $\sqcap \exists$runOn.NaturalGas $\sqcap \exists$locatedInCountry.Country

I am looking for a new car (...). This new car (should run on natural gas and should be located in a country)$_{VP}$. *N1 (V1 N1 and V2 N2)*

Example Output: Aggregation

VP Coordination

`NewCar (...) ⊓ ∃runOn.NaturalGas ⊓ ∃locatedInCountry.Country`

I am looking for a new car (...). This new car (should run on natural gas and should be located in a country)$_{VP}$. *N1 (V1 N1 and V2 N2)*

Relative Clause Coordination

`CommunicationDevice ⊓ ∃assistsWith.Understanding ⊓ ∃assistsWith.HearingDisability`

I am looking for a communication device (which should assist with a understanding and which should assist with a hearing disability)$_{RelCl}$.

Example Output: Aggregation

NP Coordination

$\texttt{CarDealer} \sqcap \exists\texttt{sell.CrashCar} \sqcap \exists\texttt{sell.NewCar}$

I am looking for a car dealer who should sell (a crash car and a new car)$_{NP}$.

Example Output: Aggregation

NP Coordination

$\texttt{CarDealer} \sqcap \exists\texttt{sell.CrashCar} \sqcap \exists\texttt{sell.NewCar}$

I am looking for a car dealer who should sell (a crash car and a new car)$_{NP}$.

N-Ary NP Coordination

$\texttt{Car} \sqcap \exists\texttt{equippedWith.ManualGearTransmission}$

$\sqcap \exists\texttt{equippedWith.AlarmSystem} \sqcap \exists\texttt{equippedWith.NavigationSystem}$

$\sqcap \exists\texttt{equippedWith.AirBagSystem}$

I am looking for a car equipped with (a manual gear transmission system, an alarm system, a navigation system and an air bag system)$_{NP}$.

Summary

- Generating from RDF Data (DBPedia, Robot tour)
- Lexicalisation (multi-triple relations, Domain-Range)
- N-ary relations
- Discourse

Summary

- Generating from RDF Data (DBPedia, Robot tour)
- Lexicalisation (multi-triple relations, Domain-Range)
- N-ary relations
- Discourse

WebNLG is looking for a Postdoc/Research Assistant/Engineer

- Machine Learning, Deep learning
- Natural Language Generation