UE903 EC1: Application to Text

Natural Language Generation (NLG)

Lecture 1: Pre-Neural NLG

Claire Gardent

Pre-Neural NLG

Generating from Data

- Some Example Data-to-Text Systems
- The D2T Pipeline

Generating from Meaning Representations

- Grammar-Based Approaches
- Statistical Models

Generating from Text

- Split, Delete, Reorder, Rewrite
- Summarisation, Compression, Paraphrasing

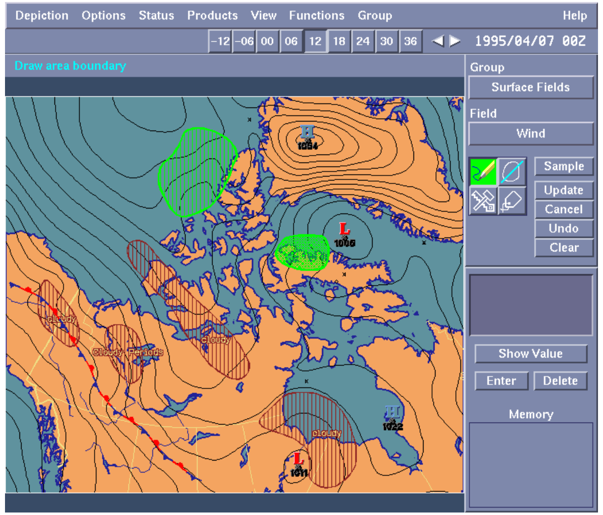
Data-to-Text Generation

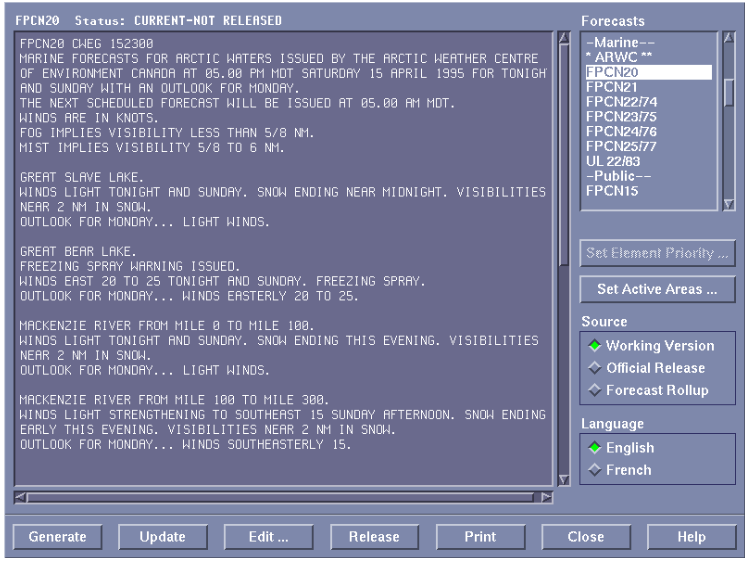
Pre-Neural D2T Generation

Applications

Example System #1: FoG

Function	Produces textual weather reports in English and French
Input	Graphical/numerical weather depiction
User	Environment Canada (Canadian Weather Service)
Developer	CoGenTex
Status	Fielded, in operational use since 1992





Pre-Neural D2T Generation

Applications

Example System #2: PlanDoc

Function	Produces a report describing the simulation options that an engineer has explored
Input	A simulation log file
User	Southwestern Bell
Developer	Bellcore and Columbia University
Status	Fielded, in operational use since 1996

RUNID FIBERALL FIBER 6/19/93 ACT
YES
FA 1301 2 1995 FA 1201 2 1995
FA 1401 2 1995 FA 1501 2 1995
ANF co 1103 2 1995 48
ANF 1201 1301 2 1995 24
ANF 1401 1501 2 1995 24
END. 856.0 670.2

This saved fiber refinement includes all DLC changes in Run-ID ALLDLC. RUN-ID FIBERALL demanded that PLAN activate fiber for CSAs 1201, 1301, 1401 and 1501 in 1995 Q2. It requested the placement of a 48-fiber cable from the CO to section 1103 and the placement of 24-fiber cables from section 1201 to section 1301 and from section 1401 to section 1501 in the second quarter of 1995. For this refinement, the resulting 20 year route PWE was $856.00K, a $64.11K savings over the BASE plan and the resulting 5 year IFC was $670.20K, a $60.55K savings over the BASE plan.

Pre-Neural D2T Generation

Applications

Example System #3: TEMSIS

Function	Summarises pollutant information for environmental officials
Input	Environmental data + a specific query
User	Regional environmental agencies in France and Germany
Developer	DFKI GmbH
Status	Prototype developed; requirements for fielded system being analysed

```
((LANGUAGE FRENCH)
(GRENZWERTLAND GERMANY)
(BESTAETIGE-MS T)
    (BESTAETIGE-SS T)
    (MESSSTATION "Voelklingen City")
    (DB-ID "#2083")
    (SCHADSTOFF "#19")
    (ART MAXIMUM)
    (ZEIT ((JAHR 1998)
           (MONAT 7)
           (TAG 21))))
```

Le 21/7/1998 à la station de mesure de Völklingen -City, la valeur moyenne maximale d'une demi-heure (Halbstundenmittelwert) pour l'ozone atteignait 104.0 µg/m³. Par conséquent, selon le decret MIK (MIK-Verordnung), la valeur limite autorisée de 120 µg/m³ n'a pas été dépassée.

Pre-Neural D2T Generation

Types of D2T Systems

Applications

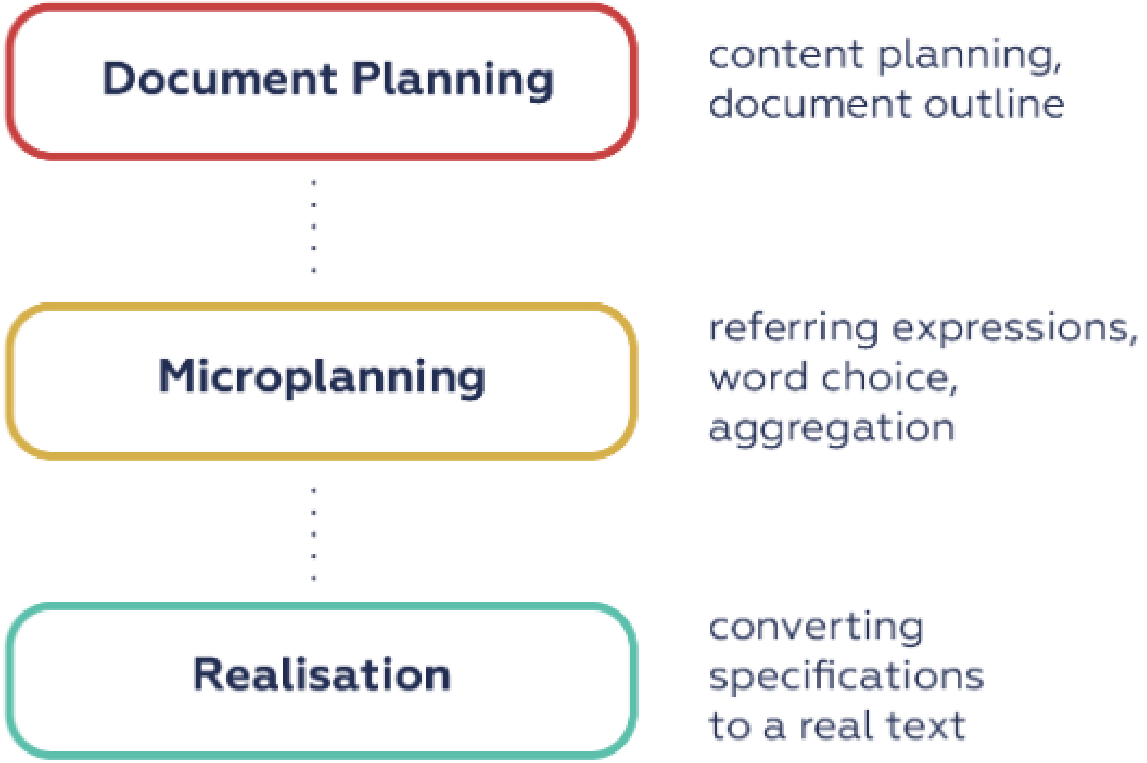
GOAL	APPLICATIONS
Automated document production	weather forecasts, simulation reports, letters, ...
Presentation of information to people in an understandable fashion	medical records, expert system reasoning, numerical data, ...
Teaching	Grammar exercises
Entertainment	Jokes, stories, poetry

Preneural D2T Generation

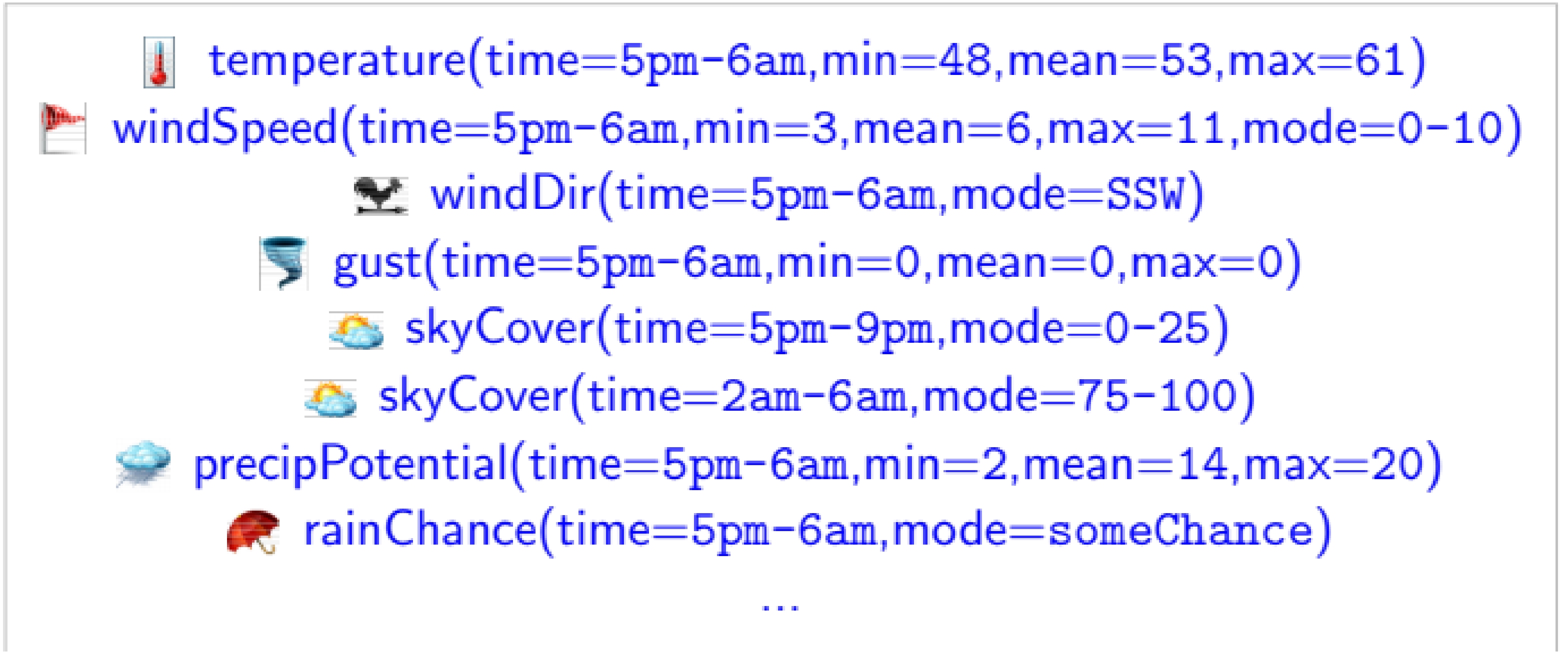
Applications

The D2T Pipeline

The D2T Pipeline



Content Selection: determining what to say



- Verbalising all the input data would yield a very redundant, unnatural text.
- Content selection selects and structures the input data so as to support the generation of a natural sounding text
- Content selection is domain specific

Preneural D2T Generation

Applications

The D2T Pipeline

Content Selection

Content Selection

Input Data

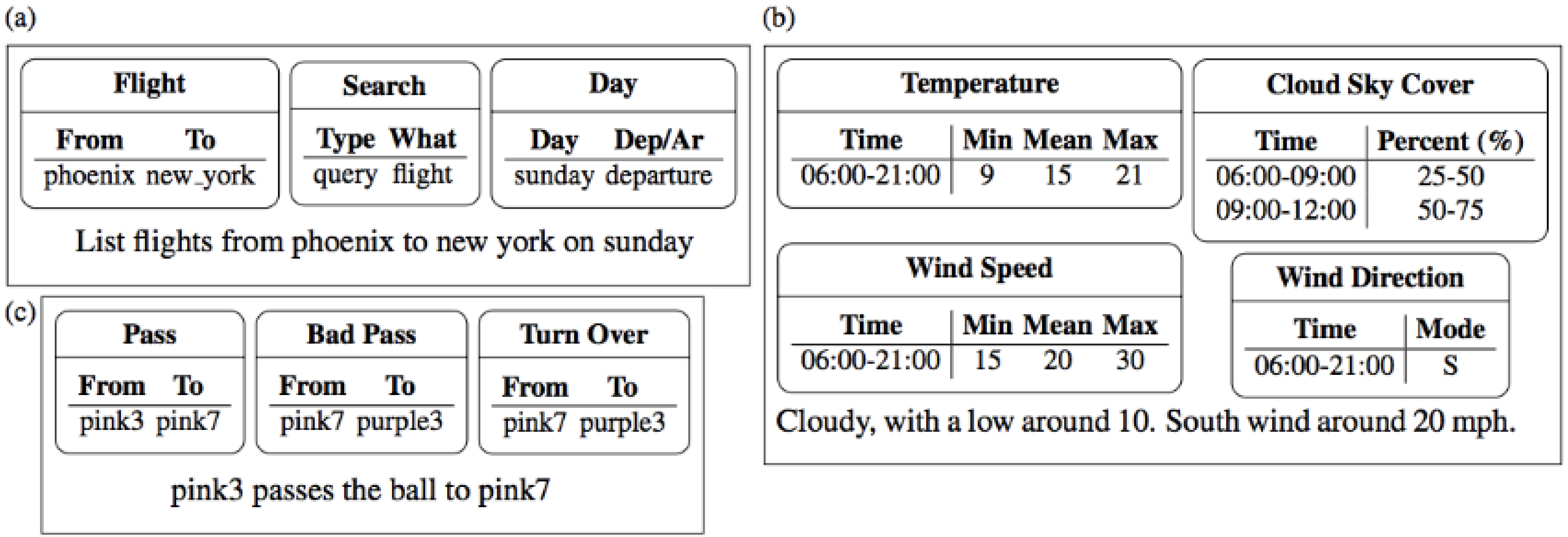
temperature(time=5pm-6am,min=48,mean=53,max=61)
windSpeed(time=5pm-6am,min=3,mean=6,max=11,mode=0-10)
windDir(time=5pm-6am,mode=SSW)
gust(time=5pm-6am,min=0,mean=0,max=0)
skyCover(time=5pm-9pm,mode=0-25)
skyCover(time=2am-6am,mode=75-100)
precipPotential(time=5pm-6am,min=2,mean=14,max=20)
rainChance(time=5pm-6am,mode=someChance)
...

Output Text

A 20 percent chance of showers after midnight. Increasing clouds, with a low around 48. Southwest wind between 5 and 10 mph.

Content Selection

Selecting content for sportscasting, weather forecast and flight booking.



Session 3: 24/09. Statistical Content Selection

ABROUGUI Rim

Gabor Angeli, Percy Liang, and Dan Klein. A simple domain-independent probabilistic approach to generation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 502--512, Cambridge, MA, October 2010. Association for Computational Linguistics. [http]

> We present a simple, robust generation system which performs content selection and surface realization in a unified, domain-independent framework. In our approach, we break up the end-to-end generation process into a sequence of local decisions, arranged hierarchically and each trained discriminatively. We deployed our system in three different domains—Robocup sportscasting, technical weather forecasts, and common weather forecasts, obtaining results comparable to state-ofthe-art domain-specific systems both in terms of BLEU scores and human evaluation.

AFARA Maria

Ioannis Konstas and Mirella Lapata. Concept-to-text generation via discriminative reranking. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 369--378, Jeju Island, Korea, July 2012. Association for Computational Linguistics. [http]

> This paper proposes a data-driven method for concept-to-text generation, the task of automatically producing textual output from non-linguistic input. A key insight in our approach is to reduce the tasks of content selection (“what to say”) and surface realization (“how to say”) into a common parsing problem. We define a probabilistic context-free grammar that describes the structure of the input (a corpus of database records and text describing some of them) and represent it compactly as a weighted hypergraph. The hypergraph structure encodes exponentially many derivations, which we rerank discriminatively using local and global features. We propose a novel decoding algorithm for finding the best scoring derivation and generating in this setting. Experimental evaluation on the ATIS domain shows that our model outperforms a competitive discriminative system both using BLEU and in a judgment elicitation study.

Pre-Neural, Symbolic Content Selection

Creates a set of messages from input data and other domain/background information

- Specific to the application domain
- Filter, summarize and process the input data
- Can be affected by a user model, history
- Can incorporate reasoning and planning algorithms

Extracting Messages from Rail Timetables

Current time: Monday 09:40

Mon-Fri	**t01**	**t02**	**t03**	**t04**	**t05**	**Caledonian Express**
Aberdeen	0533	0633	0737	0842	0937	1000
Dundee	0651	0750	0853	0952	1052	1149
Perth	0714	0812	0915	1014	1114	1211
Glasgow	0834	0915	1014	1114	1215	1314

Preneural D2T Generation

Applications

The D2T Pipeline

Content Selection

Message (Selecting)

Current time: Monday 09:40

Mon-Fri	t01	t02	t03	t04	t05	Caledonian Express
Aberdeen	0533	0633	0737	0842	0937	1000
Dundee	0651	0750	0853	0952	1052	1149
Perth	0714	0812	0915	1014	1114	1211
Glasgow	0834	0915	1014	1114	1215	1314

```
Message-id:msg02
Relation:DEPARTURE
 departing-entity:CALEDON-EXP
 departure-location:ABERDEEN
 departure-time:1000
```

Preneural D2T Generation

Applications

The D2T Pipeline

Content Selection

Message (User Adaptation)

Current time: Monday 09:40

Mon-Fri	**t01**	**t02**	**t03**	**t04**	**t05**	**Caledonian Express**
Aberdeen	0533	0633	0737	0842	0937	1000
Dundee	0651	0750	0853	0952	1052	1149
Perth	0714	0812	0915	1014	1114	1211
Glasgow	0834	0915	1014	1114	1215	1314

Next train

```
Message-id:msg01
Relation:IDENTITY
 arg1:NEXT-TRAIN
 arg2:CALEDON-EXP
```

Preneural D2T Generation

Applications

The D2T Pipeline

Content Selection

Messages (Reasoning)

Current time: Monday 09:40

Mon-Fri	t01	t02	t03	t04	t05	Caledonian Express
Aberdeen	0533	0633	0737	0842	0937	1000
Dundee	0651	0750	0853	0952	1052	1149
Perth	0714	0812	0915	1014	1114	1211
Glasgow	0834	0915	1014	1114	1215	1314

Summarising the data

```
Message-id:msg03
Relation:NUMBER-OF-TRAINS-IN-PERIOD
 traject:ABERDEEN/GLASGOW
 period:DAILY
 number:6
```

Preneural D2T Generation

Applications

The D2T Pipeline

Document Planning

Document Planning

Goal

- To determine how to structure the selected information in order to generate a coherent text

Two Common (Symbolic) Approaches

- Methods based on observations about common text structures (Schemas)
- Methods based on reasoning about discourse coherence and the purpose of the text (Discourse Structure)

Document Planning using Schemas

(McKeown 1985)

- texts often follow conventionalised patterns
- these patterns specify how a particular document plan can be constructed using smaller schemas or atomic messages

Implementing schemas

- simple schemas can be expressed as grammars
- more flexible schemas usually implemented as macros or class libraries on top of a conventional programming language, where each schema is a procedure

Preneural D2T Generation

Applications

The D2T Pipeline

Document Planning

A Simple Schema

```
WeatherSummary ->
 MonthlyTempMsg
 MonthlyRainfallMsg
 RainyDaysMsg
 RainSoFarMsg
```

A More Complex set of Schemata

- Recursive
- Includes optional elements (in square brackets)

```
WeatherSummary ->
 TemperatureInformation RainfallInformation
TemperatureInformation ->
 MonthlyTempMsg [ExtremeTempInfo] [TempSpellsInfo]
RainfallInformation ->
 MonthlyRainfallMsg [RainyDaysInfo] [RainSpellsInfo]
RainyDaysInfo ->
 RainyDaysMsg [RainSoFarMsg]
```

Schemas: Pros and Cons

Advantages of schemas

- Computationally efficient
- Allow arbitrary computation when necessary
- Naturally support genre conventions
- Relatively easy to acquire from a corpus

Disadvantages

- Limited flexibility: require predetermination of possible structures
- Limited portability: likely to be domain-specific

Preneural D2T Generation

The D2T Pipeline

Document Planning

Text Planning using Discourse Theory

Texts are coherent by virtue of relationships that hold between their parts — relationships like narrative sequence, elaboration, justification ...

Discourse-Based Content Planning

- specify discourse structure rules
- use these rules to dynamically compose texts from constituent elements by reasoning about the role of these elements in the overall text
- Typically adopt AI planning techniques
  - A text realises a plan, with a goal and subplans
  - Goal = desired communicative effect
  - Plan constituents = messages or structures that combine messages (subplans)
  - Can involve explicit reasoning about the user's beliefs
  - Often based on ideas from Rhetorical Structure Theory

Preneural D2T Generation

Applications

The D2T Pipeline

Document Planning

Example Discourse Relations

Sequence

- Two messages can be connected by a SEQUENCE relationship if both have the attribute massage-status = primary

Elaboration

- Two messages can be connected by an ELABORATION relationship if:
  - they are both have the same message-topic
  - one of the two messages has message-status = primary

Contrast

- Two messages can be connected by a CONTRAST relationship if:
  - they both have the same message-topic
  - they both have the feature absolute-or-relative = relative-to-average
  - they have different values for relative-difference:direction

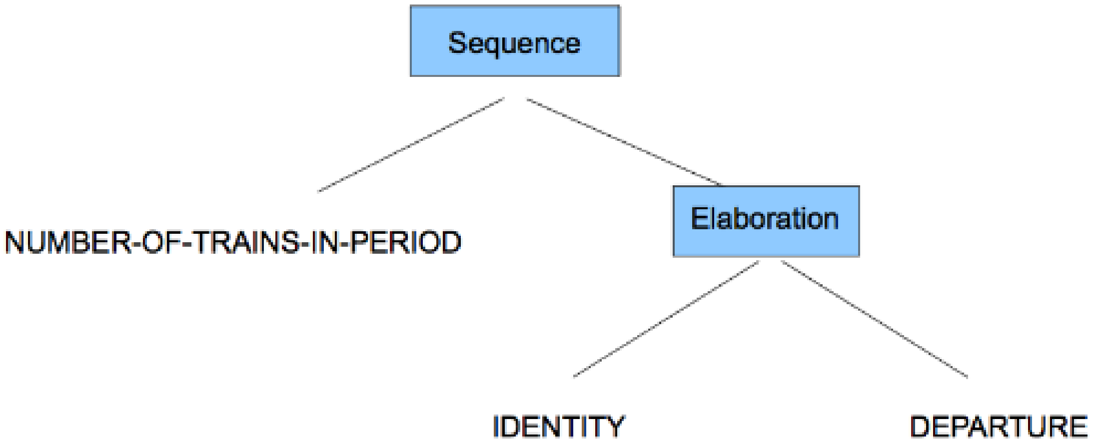
Preneural D2T Generation

Applications

The D2T Pipeline

Document Planning

Example RST Tree (Train Example)



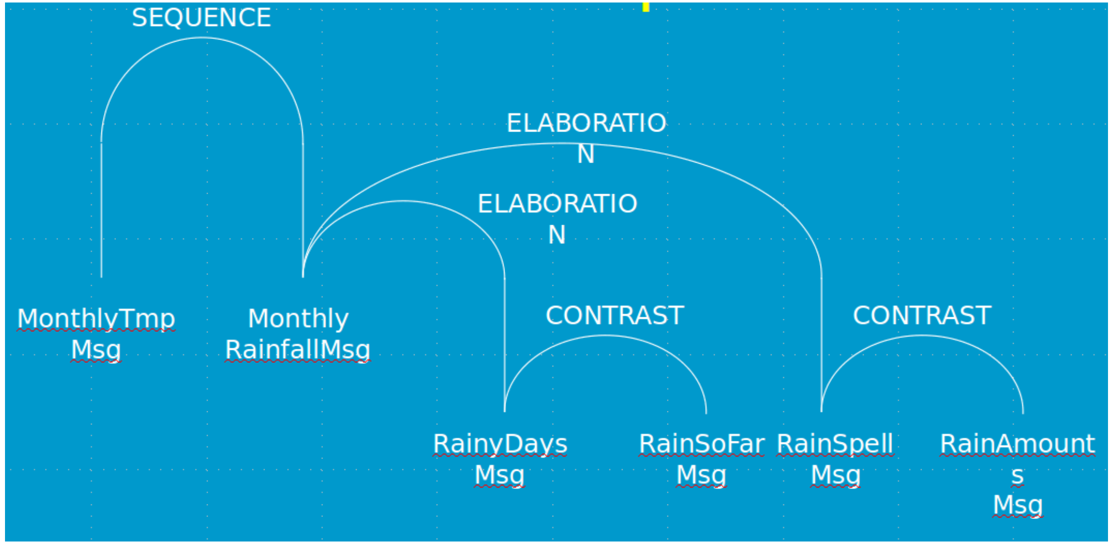
There are 6 trains a day from Aberdeen to Glasgow.
The next train is the Caledonian Express.
It departs at 10am.

Discourse Planning Algorithm

- DocumentPlan = StartMessage
- MessageSet = MessageSet - StartMessage
- repeat
    - find a rhetorical operator that will allow attachment of a message to the DocumentPlan
    - attach message and remove from MessageSet
- until MessageSet = 0 or no operators apply

Input Messages and Discourse Tree

```
MonthlyTempMsg ("cooler than average")
MonthlyRainfallMsg ("drier than average")
RainyDaysMsg ("average number of rain days")
RainSoFarMsg ("well below average")
RainSpellMsg ("8 days from 11th to 18th")
RainAmountsMsg ("amounts mostly small")
```



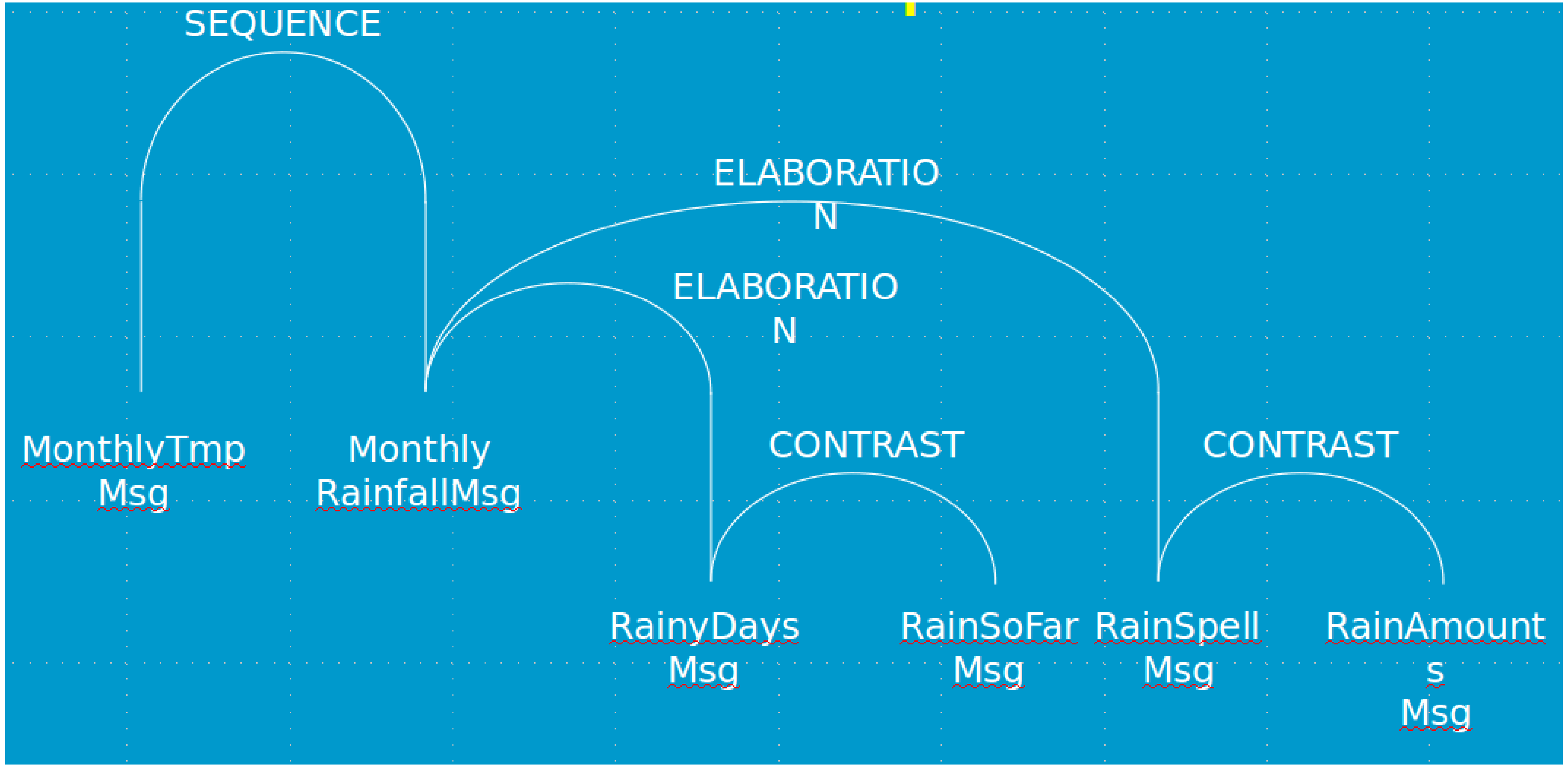
Preneural D2T Generation

Applications

The D2T Pipeline

Document Planning

Discourse Tree and Output Text



The month was cooler and drier than average, with the average number of rain days, but the total rain for the year so far is well below average. Although there was rain on every day for 8 days from 11th to 18th, rainfall amounts were mostly small.

Document Planning (Summary)

- Result = Document Plan
  - a tree structure populated by messages at its leaf nodes
- Next step: realising the messages as text

From Plan to Text

- Referring Expression Generation: Deciding how to describe entities
- Lexicalisation: Choosing words for input symbols
- Aggregation: Using Ellipsis and Coordination to avoid repetition
- Surface Realisation: Choosing the syntactic form of sentences
- Sentence segmentation: Segmenting the content into sentence size chunks

Generating Referring Expressions

Describing entities

```
(John_E_Blaha birthDate 1942_08_26)
(John_E_Blaha birthPlace San_Antonio)
(John_E_Blaha occupation  Fighter_pilot)
```

✓ *John E Blaha* was born in San Antonio on 1942-08-26. *He* worked as a fighter pilot

⊗ *John E Blaha* was born in San Antonio on 1942-08-26. *John E Blaha* worked as a fighter pilot

Session 5: 01/10: Neural REG

NADAR Fatima

Thiago Castro Ferreira, Diego Moussallem, Ákos Kádár, Sander Wubben, and Emiel Krahmer. Neuralreg: An end-to-end approach to referring expression generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, page 1959–1969, Melbourne, Australia, 2018. [http]

> Traditionally, Referring Expression Generation (REG) models first decide on the form and then on the content of references to discourse entities in text, typically relying on features such as salience and grammatical function. In this paper, we present a new approach (NeuralREG), relying on deep neural networks, which makes decisions about form and content in one go without explicit feature extraction. Using a delexicalized version of the WebNLG corpus, we show that the neural model substantially improves over two strong baselines. Data and models are publicly available.

Preneural D2T Generation

Applications

The D2T Pipeline

Microplanning

Lexicalisation

Choosing lexical items

```
(John_E_Blaha birthDate 1942_08_26)
```

- John E Blaha *was born* on 1942-08-26
- John E Blaha *'s birthdate* is 1942-08-26.

Preneural D2T Generation

Applications

The D2T Pipeline

Microplanning

Surface Realisation

Choosing syntactic structures

```
(John_E_Blaha birthPlace San_Antonio)
(John_E_Blaha birthDate 1942_08_26)
(John_E_Blaha occupation  Fighter_pilot)
```

- John E Blaha, *born in San Antonio*, on 1942-08-26 worked as a fighter pilot
- John E Blaha *was born in San Antonio* on 1942-08-26. He worked as a fighter pilot
- John E Blaha *who was born in San Antonio on 1942-08-26* worked as a fighter pilot

Preneural D2T Generation

Applications

The D2T Pipeline

Microplanning

Aggregation

Avoiding repetition

```
(John_E_Blaha birthDate 1942_08_26)
(John_E_Blaha birthPlace San_Antonio)
(John_E_Blaha occupation  Fighter_pilot)
```

- ✓ *John E Blaha* , born in San Antonio on 1942-08-26, worked as a fighter pilot
- ⊗ *John E Blaha was born* in San Antonio. *John E Blaha was born* on 1942-08-26. *John E Blaha* worked as a fighter pilot

Preneural D2T Generation

Applications

The D2T Pipeline

Microplanning

Sentence segmentation

Segmenting the content into sentence size chunks

```
(John_E_Blaha birthDate 1942_08_26)
(John_E_Blaha birthPlace San_Antonio)
(John_E_Blaha occupation  Fighter_pilot)
```

- John E Blaha, born in San Antonio on 1942-08-26, worked as a fighter pilot.
- John E Blaha was born in San Antonio on 1942-08-26. He worked as a fighter pilot.

MR-to-Text Generation

Pre-Neural NLG

D2T

MR2T

Generating from Meaning Representations

Some Example Tasks

Grammar Based Approaches

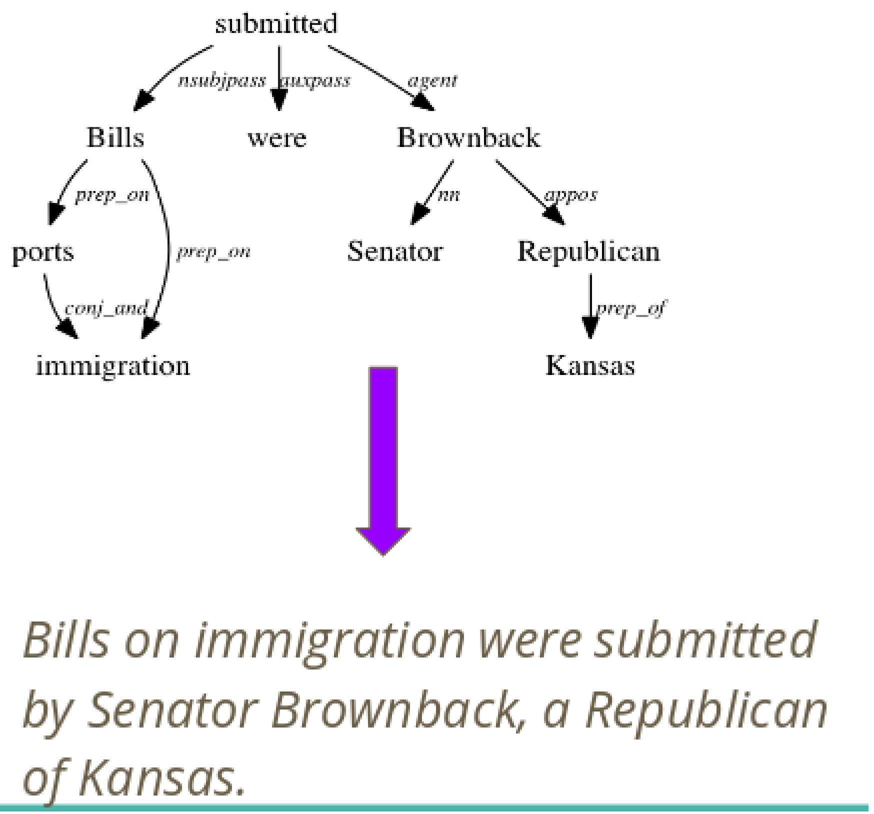
Statistical Models

MR2T

Example Tasks

Example #1: Dependency Trees

Surface Realization Challenge 2011 and 2018



MR2T

Example Tasks

Example #2: Abstract Meaning Representations (AMR)

```
(s / say-01
      :ARG0 (s2 / service
            :mod (e / emergency)
            :location (c / city :wiki ''London''
                  :name (n / name :op1 ''London'')))
      :ARG1 (s3 / send-01
            :ARG1 (p / person :quant 11)
            :ARG2 (h / hospital)
            :mod (a / altogether)
            :purpose (t / treat-03
                  :ARG1 p
                  :ARG2 (w / wound-01
                        :ARG1 p
                        :mod (m / minor)))))
```

The London emergency services said that altogether 11 people had been sent to hospital for treatment due to minor wounds.

MR2T

Example Tasks

Example System #3: Dialog Moves

Recommend
name[The Eagle],
eatType[coffee shop],
food[French],
priceRange[moderate],
customerRating[3/5],
area[riverside],
kidsFriendly[yes],
near[Burger King]

The three star coffee shop, The Eagle, gives families a mid-priced dining experience featuring a variety of wines and cheeses. Find The Eagle near Burger King

MR2T

Example Tasks

Grammar-Based Approaches

Generating from Description Logic Formulae

The input Query is converted to a flat representation

DL: Professor $\sqcap$ Researcher $\sqcap$ $\exists$teach.LogicCourse $\sqcap$ $\exists$worksAt.AlicanteUniversity

Flat: Professor(p) Researcher(p) teach(p c) LogicCourse(c) worksAt(p u) AlicanteUniversity(u)

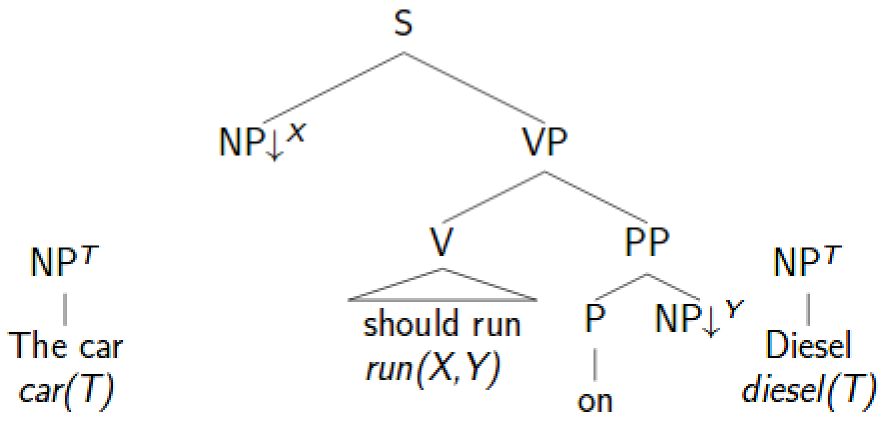
MR2T

Example Tasks

Grammar-Based Approaches

Tree Adjoining Grammar

- The grammar is a set of trees (rules)
- Grammar rules (trees) map words to syntactic trees and semantic representations (flat representation of DL subformulae)
- Trees can be combined using substitution (↓) or adjunction (*)



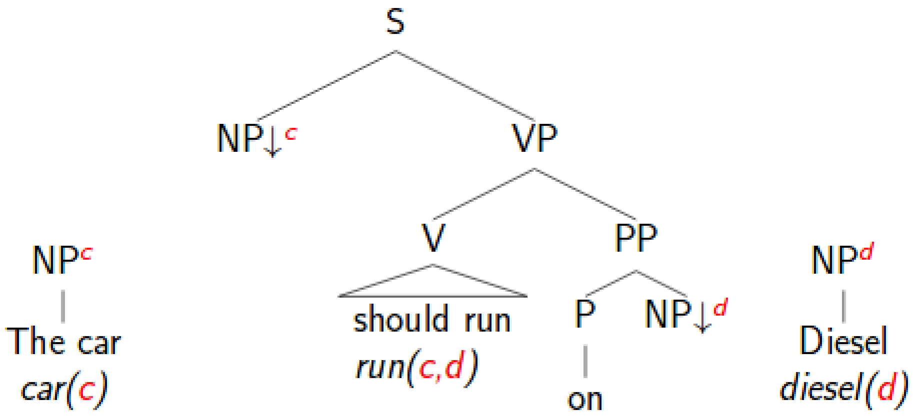
MR2T

Example Tasks

Grammar-Based Approaches

Step 1: Lexical Selection

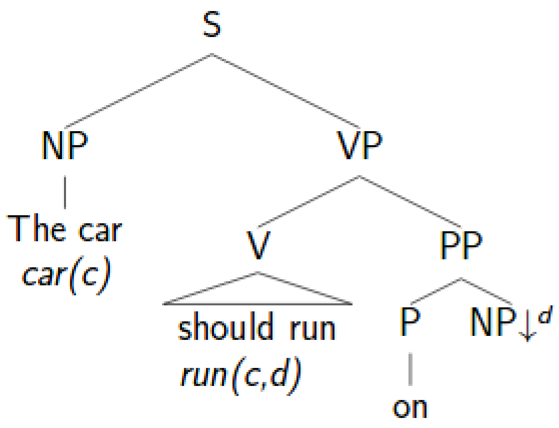
Trees whose semantics subsume the input are selected (variables are unified)



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

Step 2: Combining trees

The tree for "The car" is substituted in the "should run on" tree



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

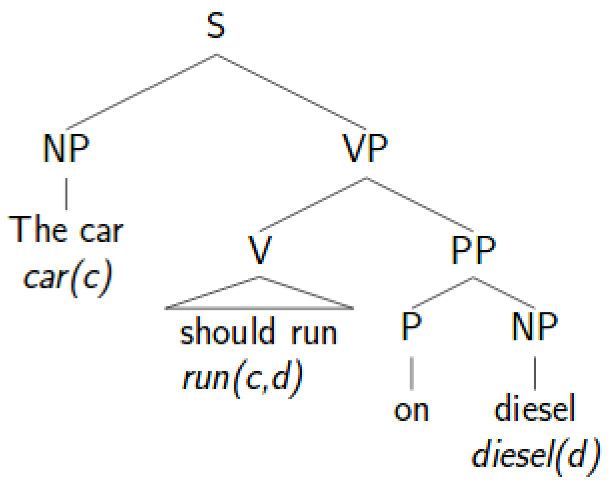
MR2T

Example Tasks

Grammar-Based Approaches

Step 2: Combining trees

The tree for "diesel" is substituted in the "the car should run on" tree



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

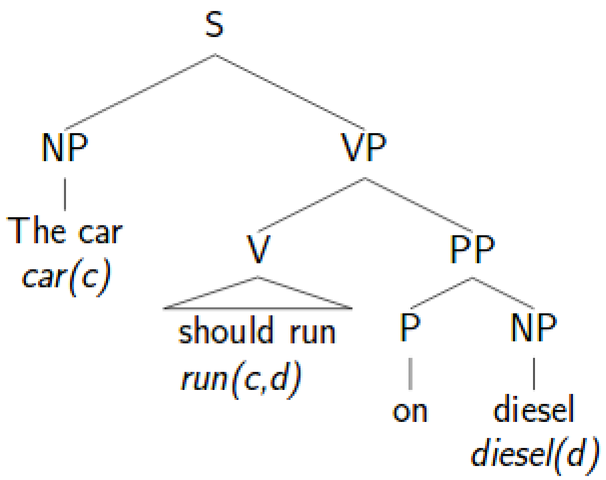
MR2T

Example Tasks

Grammar-Based Approaches

Step 3: Extraction

Sentences are generated by extracting the yield of all trees which are syntactically complete and which cover the input semantics




```
The car should run on diesel
```

MR2T

Example Tasks

Grammar-Based Approaches

Hybrid Statistical, Grammar-Based Approaches

- Grammar-Based approach yields multiple outputs and intermediate results
- Statistical modules are used to reduce ambiguity
    - Language models
      To choose between comparable intermediate results (the black cat/the cat black)
    - Hypertaggers
      To prune the initial search space
    - Rankers
      To determine the best output

MR2T

Example Tasks

Grammar-Based Approaches

Statistical Approaches

Overgenerate and Rank

Langkilde 1998

- Generates from AMR
- A lexicon and keyword based grammar rules are used to map the input AMR to words
- A Language Model is used to extract from the resulting lattice the sentence with highest probability.

Cascaded Classifiers

Bohnet et al. 2010

- Generates from Multilevel Annotated (Parallel) Data
- Each Level encodes a different way of representing the output text (from deeper to more surface level representations
- For each pair of adjacent levels of annotation, a separate SVM (Support Vector Machine) decoder is learned.

Session 3: 24/09. Statistical MR2T Generation

AKANI Aduenu

Irene Langkilde and Kevin Knight. Generation that exploits corpus-based statistical knowledge. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*, 1998. [http]

> We present a simple, robust generation system which performs content selection and surface realization in a unified, domain-independent framework. In our approach, we break up the end-to-end generation process into a sequence of local decisions, arranged hierarchically and each trained discriminatively. We deployed our system in three different domains—Robocup sportscasting, technical weather forecasts, and common weather forecasts, obtaining results comparable to state-ofthe-art domain-specific systems both in terms of BLEU scores and human evaluation.

AKHMETOV Alisher

Bernd Bohnet, Leo Wanner, Simon Mille, and Alicia Burga. Broad coverage multilingual deep sentence generation with a stochastic multi-level realizer. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 98--106, Beijing, China, August 2010. Coling 2010 Organizing Committee. [http]

> Most of the known stochastic sentence generators use syntactically annotated corpora, performing the projection to the surface in one stage. However, in full-fledged text generation, sentence realization usually starts from semantic (predicate-argument) structures. To be able to deal with semantic structures, stochastic generators require semantically annotated, or, even better, multilevel annotated corpora. Only then can they deal with such crucial generation issues as sentence planning, linearization and morphologization. Multilevel annotated corpora are increasingly available for multiple languages. We take advantage of them and propose a multilingual deep stochastic sentence realizer that mirrors the state-ofthe-art research in semantic parsing. The realizer uses an SVM learning algorithm. For each pair of adjacent levels of annotation, a separate decoder is defined. So far, we evaluated the realizer for Chinese, English, German, and Spanish.

Text-to-Text Generation

Pre-Neural NLG

D2T

MR2T

T2T

Main Tasks

Four Base Operations

Approaches

T2T

Main Tasks

Example #1: Summarization

REVIEW

doi:10.1038/nature14539

Deep learning

Yann LeCun[1,2], Yoshua Bengio[3] & Geoffrey Hinton[4,5]

Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction. These methods have dramatically improved the state-of-the-art in speech recognition, visual object recognition, object detection and many other domains such as drug discovery and genomics. Deep learning discovers intricate structure in large data sets by using the backpropagation algorithm to indicate how a machine should change its internal parameters that are used to compute the representation in each layer from the representation in the previous layer. Deep convolutional nets have brought about breakthroughs in processing images, video, speech and audio, whereas recurrent nets have shone light on sequential data such as text and speech.

Machine-learning technology powers many aspects of modern society: from web searches to content filtering on social networks to recommendations on e-commerce websites, and it is increasingly present in consumer products such as cameras and smartphones. Machine-learning systems are used to identify objects in images, transcribe speech into text, match news items, posts or products with users' interests, and select relevant results of search. Increasingly, these applications make use of a class of techniques called deep learning.

Conventional machine-learning techniques were limited in their ability to process natural data in their raw form. For decades, constructing a pattern-recognition or machine-learning system required careful engineering and considerable domain expertise to design a feature extractor that transformed the raw data (such as the pixel values of an image) into a suitable internal representation or feature vector from which the learning subsystem, often a classifier, could detect or classify patterns in the input.

intricate structures in high-dimensional data and is therefore applicable to many domains of science, business and government. In addition to beating records in image recognition[1–4] and speech recognition[5–7], it has beaten other machine-learning techniques at predicting the activity of potential drug molecules[8], analysing particle accelerator data[9,10], reconstructing brain circuits[11], and predicting the effects of mutations in non-coding DNA on gene expression and disease[12,13]. Perhaps more surprisingly, deep learning has produced extremely promising results for various tasks in natural language understanding[14], particularly topic classification, sentiment analysis, question answering[15] and language translation[16,17].

We think that deep learning will have many more successes in the near future because it requires very little engineering by hand, so it can easily take advantage of increases in the amount of available computation and data. New learning algorithms and architectures that are currently being developed for deep neural networks will only accelerate this progress.

Grusky et al., 2018

T2T

Main Tasks

Example #2: Simplification

Complex

In 1964 Peter Higgs published his second paper in Physical Review Letters describing Higgs mechanism which predicted a new massive spin-zero boson for the first time.

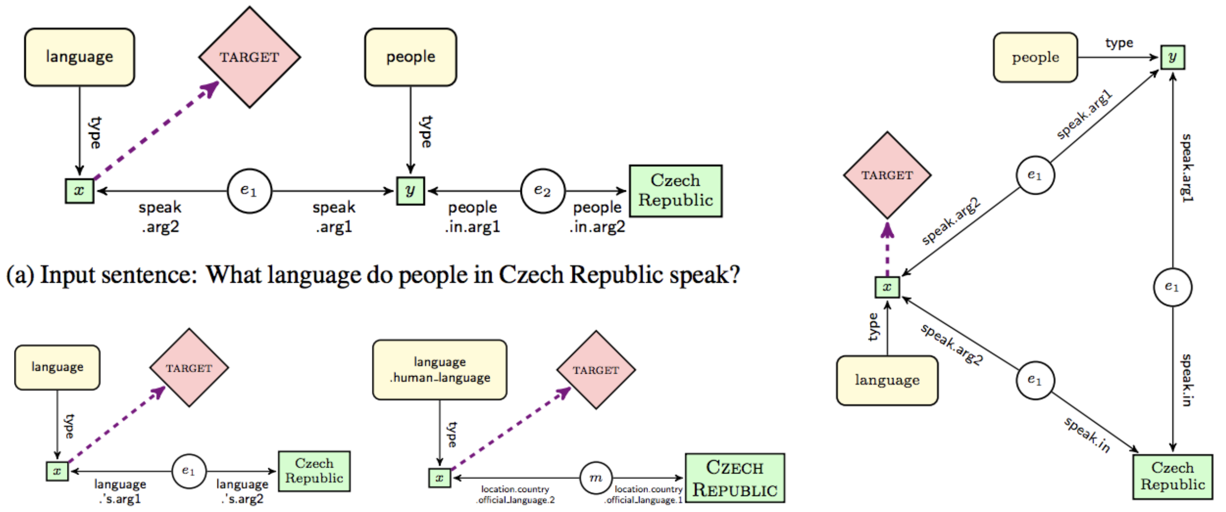
Simple

Peter Higgs wrote his paper explaining Higgs mechanism in 1964. Higgs mechanism predicted a new elementary particle.

T2T

Main Tasks

Example #3: Paraphrasing



(a) Input sentence: What language do people in Czech Republic speak?

(c) Paraphrase: What is Czech Republic's language?

(d) Freebase grounded graph

(b) Paraphrase: What language do people speak in Czech Republic?

Narayan et al, 2016

T2T

Main Tasks

Operations

Split

Rewrite

Move

Delete

Split

In 1964 Peter Higgs published his ~~second~~ paper ~~in Physical Review Letters~~ describing Higgs mechanism [which] predicted a new massive spin-zero boson ~~for the first time~~.

SPLIT

Peter Higgs wrote his paper explaining Higgs mechanism in 1964.

Higgs mechanism predicted a new elementary particle.

T2T

Main Tasks

Operations

Move

In 1964 Peter Higgs published his ~~second~~ paper ~~in Physical Review Letters~~ describing Higgs mechanism [which] predicted a new massive spin-zero boson ~~for the first time~~.

REORDER

Peter Higgs wrote his paper explaining Higgs mechanism in 1964.

Higgs mechanism predicted a new elementary particle.

T2T

Main Tasks

Operations

Rewrite

In 1964 Peter Higgs published his ~~second~~ paper ~~in Physical Review Letters~~ describing Higgs mechanism [which] predicted a new massive spin-zero boson ~~for the first time~~.

REWRITE

Peter Higgs wrote his paper explaining Higgs mechanism in 1964.

Higgs mechanism predicted a new elementary particle.

Delete

In 1964 Peter Higgs published his ~~second~~ paper ~~in Physical Review Letters~~ describing Higgs mechanism [which] predicted ~~a~~ new massive spin-zero boson ~~for the first time~~.

DELETE

Peter Higgs wrote his paper explaining Higgs mechanism in 1964.

Higgs mechanism predicted a new elementary particle.

T2T

Main Tasks

Operations

Approaches

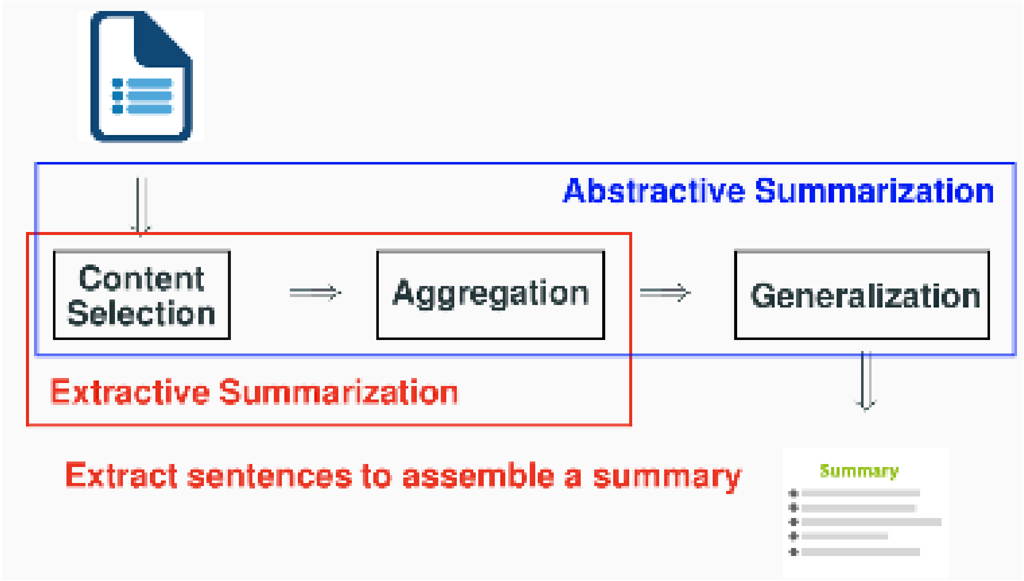
Rule-Based

- Manually specify rules for splitting, reordering, deleting and splitting

Statistical

- Learn the four operations
- Training Data: Simple and Normal Wikipedia

Abstractive vs. Extractive Summarization



- Human summarization (ABSTRACTIVE)
  - extract key information from input document
  - aggregate this information (remove redundancies)
  - abstract (produce coherent text)
- Automatic text summarization (mostly EXTRACTIVE)
  - extract key sentences from the input document
  - combine them to form a summary.

T2T

Main Tasks

Operations

Approaches

Summarization

Constraints on Extractive Summary

The generated summary

- should not exceed a given **length**
- should contain all **relevant information** and
- should **avoid repetitions**.

Three Main Steps

- create an intermediate representation of the input sentences
- score these sentences based on that representation
- create a summary by selecting the highest scoring sentences

T2T

Main Tasks

Operations

Approaches

Summarization

Three main ways of representing text

- Frequency-based
    - Probability
    - LLR
    - TF*IDF
    - Graph
- Semantics
    - Lexical chains
    - Latent Semantic Analysis
    - Discourse Structure
- Vectorial

T2T

Main Tasks

Operations

Approaches

Summarization

Lexical Frequency (Luhn 1958)

- In a document, words that are frequent are descriptive of the document's content
- The sentences that convey the most important information in a document are those that contain many such descriptive words.
- Extract important sentences

Frequency approaches

- use word probability to identify those words that represent a document topic and a frequency threshold to identify frequent content (non stop) words in a document as descriptive of the document's topic.
- (Conroy et al. 2006) used the log-likelihood ratio test to extract those words that have a likelihood statistic greater than what one would expect by chance.
- Other approaches have used TF.IDF ratio

Session 4: 26/09. Statistical and Symbolic T2T Generation

BALARD Srilakshmi

Advaith Siddharthan. Text Simplification using Typed Dependencies: A Comparison of the Robustness of Different Generation Strategies. In *Proceedings of the 13th European Workshop on Natural Language Generation (ENLG),*, pages 2--, Nancy, France, September 2011. Association for Computational Linguistics. [http]

> We present a framework for text simplification based on applying transformation rules to a typed dependency representation produced by the Stanford parser. We test two approaches to regeneration from typed dependencies: (a) gen-light, where the transformed dependency graphs are linearised using the word order and morphology of the original sentence, with any changes coded into the transformation rules, and (b) gen-heavy, where the Stanford dependencies are reduced to a DSyntS representation and sentences are generating formally using the RealPro surface realiser. The main contribution of this paper is to compare the robustness of these approaches in the presence of parsing errors, using both a single parse and an n-best parse setting in an overgenerate and rank approach. We find that the gen-light approach is robust to parser error, particularly in the n-best parse setting. On the other hand, parsing errors cause the realiser in the genheavy approach to order words and phrases in ways that are disliked by our evaluators.

BOUZIGUES Aymeric

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1353--1361. Coling 2010 Organizing Committee, 2010. [http]

> In this paper, we consider sentence simplification as a special form of translation with the complex sentence as the source and the simple sentence as the target. We propose a Tree-based Simplification Model (TSM), which, to our knowledge, is the first statistical simplification model covering splitting, dropping, reordering and substitution integrally. We also describe an efficient method to train our model with a large-scale parallel dataset obtained from the Wikipedia and Simple Wikipedia. The evaluation shows that our model achieves better readability scores than a set of baseline systems.

Session 4: 26/09. Statistical and Symbolic T2T Generation

DIEUDONAT Lea

JohnM. Conroy, JudithD. Schlesinger, and DianneP. O'Leary. Topic-focused multi-document summarization using an approximate oracle score. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 152--159, Sydney, Australia, July 2006. Association for Computational Linguistics. [http]

> We consider the problem of producing a multi-document summary given a collection of documents. Since most successful methods of multi-document summarization are still largely extractive, in this paper, we explore just how well an extractive method can perform. We introduce an “oracle” score, based on the probability distribution of unigrams in human summaries. We then demonstrate that with the oracle score, we can generate extracts which score, on average, better than the human summaries, when evaluated with ROUGE. In addition, we introduce an approximation to the oracle score which produces a system with the best known performance for the 2005 Document Understanding Conference (DUC) evaluation.

Session 4: 26/09. Evaluation

GUILLAUME Maxime

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311--318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. [bib | DOI | http]

> Human evaluations of machine translation are extensive but expensive. Human evaluations can take months to finish and involve human labor that can not be reused. We propose a method of automatic machine translation evaluation that is quick, inexpensive, and language-independent, that correlates highly with human evaluation, and that has little marginal cost per run. We present this method as an automated understudy to skilled human judges which substitutes for them when there is need for quick or frequent evaluations.

Session 5: 01/10. Evaluation

HAN Kelvin

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74--81, Barcelona, Spain, July 2004. Association for Computational Linguistics. [bib | http]

> ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation. It includes measures to automatically determine the quality of a summary by comparing it to other (ideal) summaries created by humans. The measures count the number of overlapping units such as n-gram, word sequences, and word pairs between the computer-generated summary to be evaluated and the ideal summaries created by humans. This paper introduces four different ROUGE measures: ROUGE-N, ROUGE-L, ROUGE-W, and ROUGE-S included in the ROUGE summarization evaluation package and their evaluations. Three of them have been used in the Document Understanding Conference (DUC) 2004, a large-scale summarization evaluation sponsored by NIST.