AN INTRODUCTION TO CONTENT DETERMINATION

Gerard Casamayor

Chris Mellish

Contents

1. The place of Content Determination
2. Styles of Content Determination
3. Methods for Content Determination
4. Examples
5. Content Determination from SW data

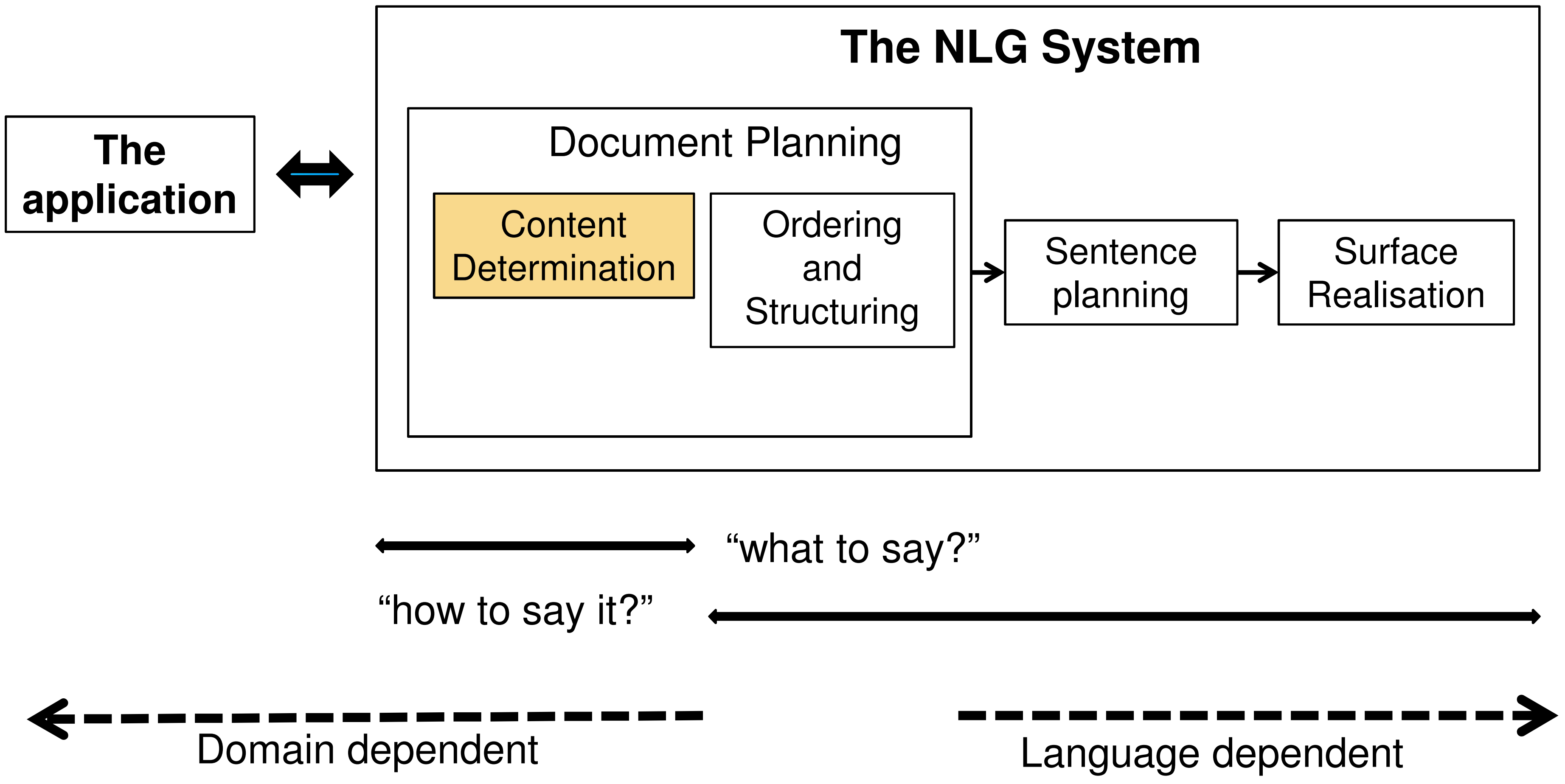
1. The place of Content Determination

Content Determination

- The main interface between the NLG system and the domain/application/outside world.

- Decides “what to say” in terms of domain concepts

Reiter and Dale (2000) NLG pipeline


The application

The NLG System

Document Planning

Content Determination

Ordering and Structuring

Sentence planning

Surface Realisation

"what to say?"

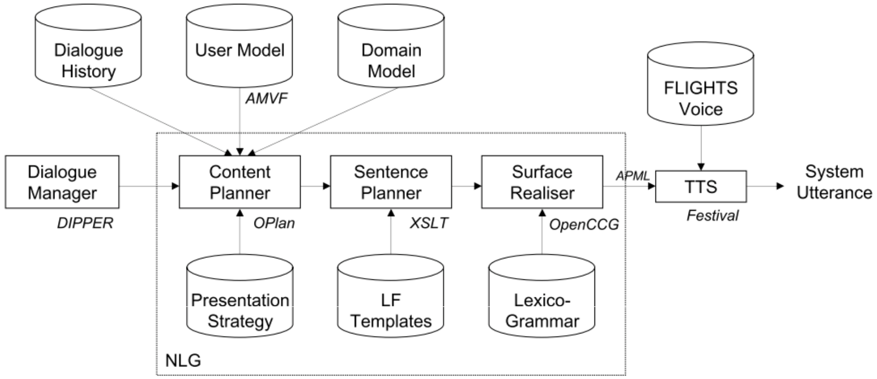
"how to say it?"

Domain dependent

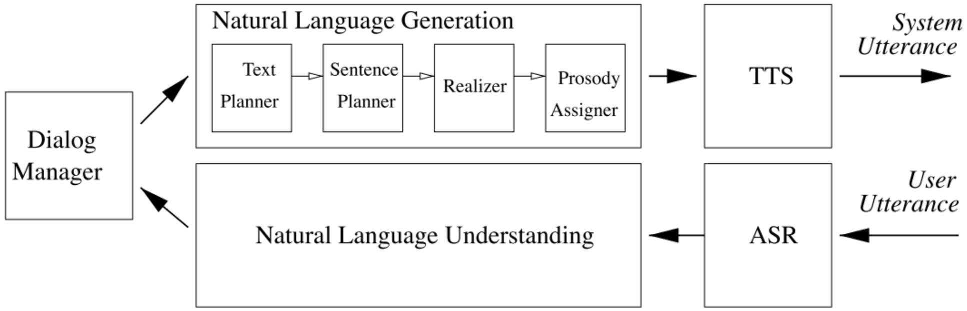
Language dependent

NLG pipelines in dialogue systems

White, Clark, and Moore (2010)



Walker, Rambow, and Rogati (2002)



- See slides on Statistical Natural Language Generation by M.White (2010) http://winterfest.hcsnet.edu.au/files2/2010/winterfest/white-bowral-part1v2.pdf

Why is Content Determination hard?

a) Hard to develop reusable approaches:

- Multiple domains
- Multiple input data formats

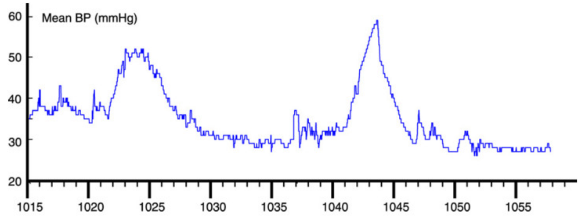
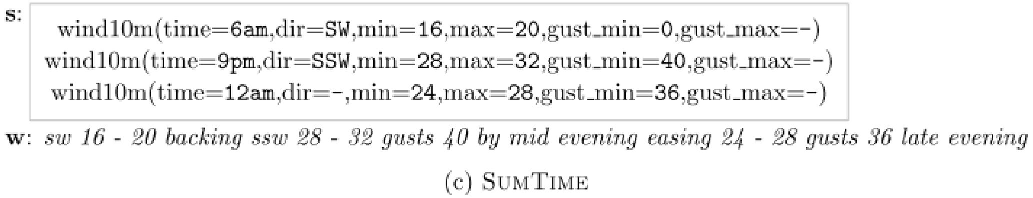


Fig. 15. Blood Pressure in scenario 14.

Continuous signal, e.g. BabyTalk (Portet et al. 2007)



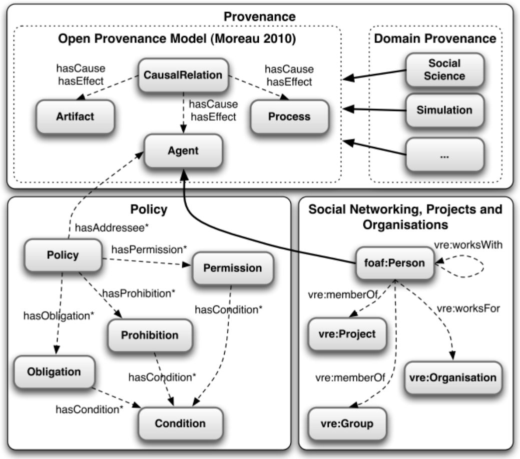
s:

```
wind10m(time=6am,dir=SW,min=16,max=20,gust_min=0,gust_max=-)
wind10m(time=9pm,dir=SSW,min=28,max=32,gust_min=40,gust_max=-)
wind10m(time=12am,dir=-,min=24,max=28,gust_min=36,gust_max=-)
```

w: *sw 16 - 20 backing ssw 28 - 32 gusts 40 by mid evening easing 24 - 28 gusts 36 late evening*

(c) SumTime

Tabular (Angeli et al. 2010)



Semantic data (Bouttaz et al. 2011)

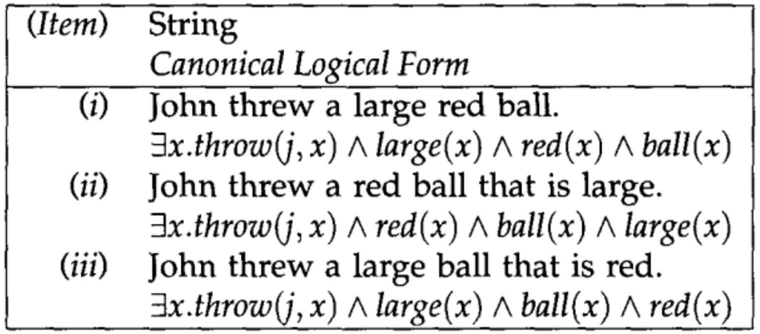
Why is Content Determination hard?

a) It may not naturally provide enough information to satisfy what the language needs, or it may not produce something that can be elegantly expressed – the “generation gap” (Meteer 92), e.g.

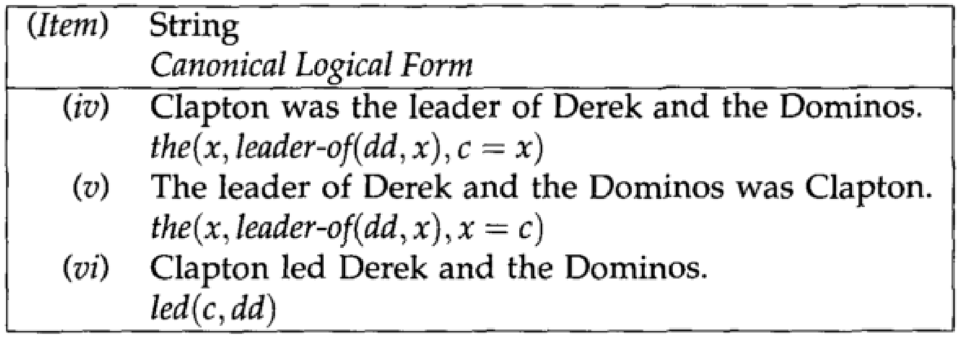
- How much material can be put into a single sentence/ paragraph/ tweet/ A4 page?
- Is it easy to express “pleasure in another person’s misfortune” (yes, if you are speaking German)?

Why is Content Determination hard?

c) It may not be able to choose among alternatives which are equivalent in the application but which make a big difference in the language, e.g. the “problem of logical form equivalence” (Shieber 93):



(Item)	String *Canonical Logical Form*
(i)	John threw a large red ball. $\exists x.throw(j,x) \land large(x) \land red(x) \land ball(x)$
(ii)	John threw a red ball that is large. $\exists x.throw(j,x) \land red(x) \land ball(x) \land large(x)$
(iii)	John threw a large ball that is red. $\exists x.throw(j,x) \land large(x) \land ball(x) \land red(x)$



(Item)	String *Canonical Logical Form*
(iv)	Clapton was the leader of Derek and the Dominos. $the(x, leader\text{-}of(dd,x), c = x)$
(v)	The leader of Derek and the Dominos was Clapton. $the(x, leader\text{-}of(dd,x), x = c)$
(vi)	Clapton led Derek and the Dominos. $led(c, dd)$

2. Styles of Content Determination

Top-down vs Bottom-up

- Top-down (goal driven, backwards) processing looks at how to find content to support one of a known set of possible text types:
  - Satisfy communicative goals
  - Good when there are strong conventions for what texts should be like
  - Making sure the text will have a coherent structure

- Bottom-up (data driven, forwards) processing looks at what the application makes available and seeing how a text can be made from it:
  - Diffuse goals
  - Working out what is most important/interesting
  - Good when the form of the text needs to vary a lot according to what is actually there

Separate task vs interleaved

- Reiter and Dale's pipeline shows Content Determination as a separate module.
- But there are dependencies between CD and other NLG tasks.
  - Error propagation: the generation gap may become evident during surface realization.
  - Alternative architectures attempt to capture interdependencies:
    - NLG systems as a unified planning problem, e.g. (Hovy 1993), (Young and Moore 1994)
    - Cascade of classifiers in the Discrete Optimization Model of (Marciniak, Strube 2005)
    - Hierarchical Reinforcement Learning for Adaptive Text Generation (Dethlefs et al. 2010)

Types of input data

- Many types of input data
- Input contents may require interpretation:

1. Continuous data signal or raw numerical data requires assessment
   - E.g. infer qualitative rating *Strong* from quantitative wind speed readings SUMTIME (Sripada et al. 2003)
2. Some aspects of the input data not explicitly encoded but inferable.
   - E.g. football match score is 1-1. Infer this result is a draw. (Bouayad-Agha et al. 2011)
3. What are the units to be selected? What is the granularity of content determination?
   - Message determination
     - In relation databases: a single cell, a whole row, a subset of the row?
     - In Semantic Web datasets: a triple, all triples about an individual?

Context

- Content determination may take into account some of the following:
  - Targeted genre: term definition, report, commentary, narrative, etc.
  - Targeted audience: lay person, informed user, domain expert, etc.
  - Request: information solicitation, decision support request, etc.
  - Communicative goal: exhaustive information on a theme, advice, persuasion, etc.
  - User profile: user preferences, needs or interests in the topic, individual expertise, previous knowledge, discourse history, etc.

3. Methods for Content Determination

Templates and schemas

- Simple and effective way of capturing observed regularities in target texts
- Templates lack flexibility
- Schemas make up for that by introducing expansion slots to be completed with contents or linguistic information.
  - (McKeown 1992)
  - MIAKT and ONTOSUM systems, (Bontcheva and Wilks 2004), (Bontcheva 2005)
- Templates and schemas can be used to by-pass NLG altogether

Automated planning

- Find sequence of actions to satisfy a goal
- Knowledge about domain and how to communicate it is modeled using planning languages (STRIPS, ADL, PDDL).
- The planning problem is addressed using a general problem solver, e.g. hierarchical planning with goal decomposition.
  - (Hovy 1993), (Young and Moore 1994), (Carenini and Moore 2006), (Paris et al. 2010).
- Content determination and structuring (and even other NLG tasks!) are handled together.
  - Planning guided by rhetorical operators that ensure coherence of text

Automated reasoning

- Start from a Knowledge Base (KB) encoding knowledge about the domain.
  - Rich semantics: knowledge representation languages, ontologies
  - Not created specifically for NLG purposes
- Types of knowledge (Rambow 1990):
  - Domain knowledge: input data, its syntax and semantics
  - Communication knowledge: domain-independent knowledge about language, discourse, etc.
  - Domain communication knowledge: how to communicate domain data
- Reasoning requires explicit, symbolic representations of how to communicate data (rules, ontologies, etc.) or a special type of inference suitable for NLG.
  - Donnell et al. (2001), (Bouayad-Agha et al. 2011, 2012), (Bouttaz 2011)
  - (Mellish and Pan 2008)

Graph-based methods

- Build a graph representation of the input data and operate on this representation.
- Graph may be reflect semantic relations between data but also statistical information, e.g. using weights.
- Two mechanisms:

  1. Explore the graph from a central point, e.g. entity of interest.
     - In Donnell et al. (2001) and Dannélls et al. (2009), a rooted content graph is navigated in search of relevant data.
  2. Apply a global graph algorithm to weight all nodes/Edges and find most relevant subset.
     - In Demir et al. (2010) PageRank is applied to find a subset of the content graph that maximizes relevance and reduces redundancies.

Statistical methods

- General statistical approach:
  1. Construct a general model that assigns probabilities to outputs, given inputs
  2. Provide training data to the model, in order to tune the internal parameters
  3. Present the trained model with a real input
  4. Search for the output which maximises the probability according to the model

- Model can be trained from corpora of human authored texts aligned with contents.
  a. Manual annotation
  b. Automatic linkage of texts and contents

Statistical methods

System	Model	Input	Search strategy	Training data
Barzilay and Lapata 2005	Weighted graph + multiple classifiers	Database rows	Minimal cut partition	Automatically aligned corpus
Kelly et al. 2009	Single classifier	Semistructured data	None	Automatically aligned corpus
Belz 2008	PCFG with estimated weights	Tabular	Greedy	Manually annotated corpus
Konstas et a. 2013	PCFG with estimated weights	Database rows	CYK	Manually annotated corpus
Rieser et al. 2010	Markov Decision Process	Database cells	Reinforcement Learning	Feedback from simulated user
Dethlefs et al. 2011	Markov Decision Process	Simulated data	Hierarchical reinforcement learning	Feedback from simulated user

Wrapping up

- Styles:
  1. Top-down vs bottom-up
  2. Separate task vs interleaved
  3. Type of input data
  4. Context
- Methods:
  1. Templates and schemas
  2. Automated planning
  3. Automated reasoning
  4. Graph-based methods
  5. Statistical methods
- Methods aren’t mutually exclusive, they can be combined in the same implementation.

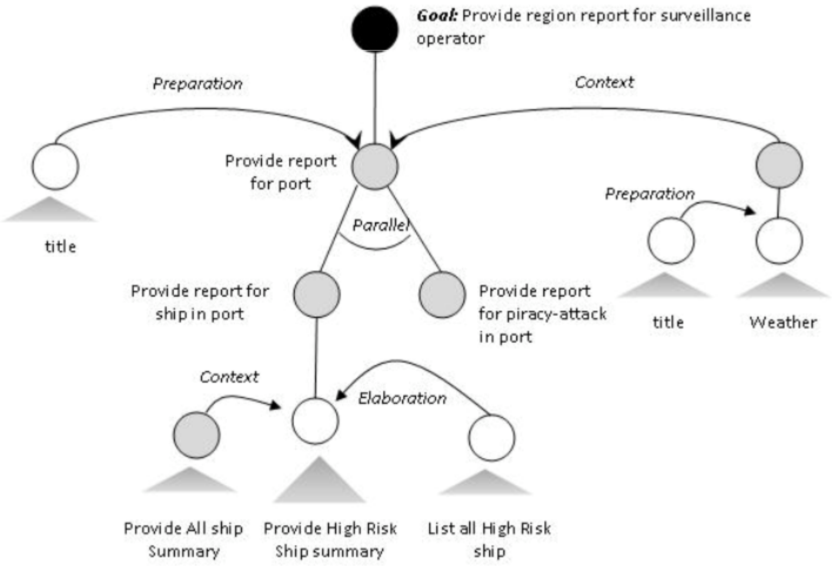
4. Examples

Example 1: Paris et al. 2010

- Input data:
  - Knowledge Base with explicit semantics (domain ontology).
  - Granularity: coarse-grained units of information.
- Top-down, goal-driven.
- Interleaved with structuring
- Context: user profile, user history, explicit communicative goals
- Methods: hierarchical planning.

Example 1: Paris et al. 2010

- Text planning module produces discourse structures where information is connected with rhetorical relations.
  - Discourse trees from Rhetorical Structure Theory (RST).

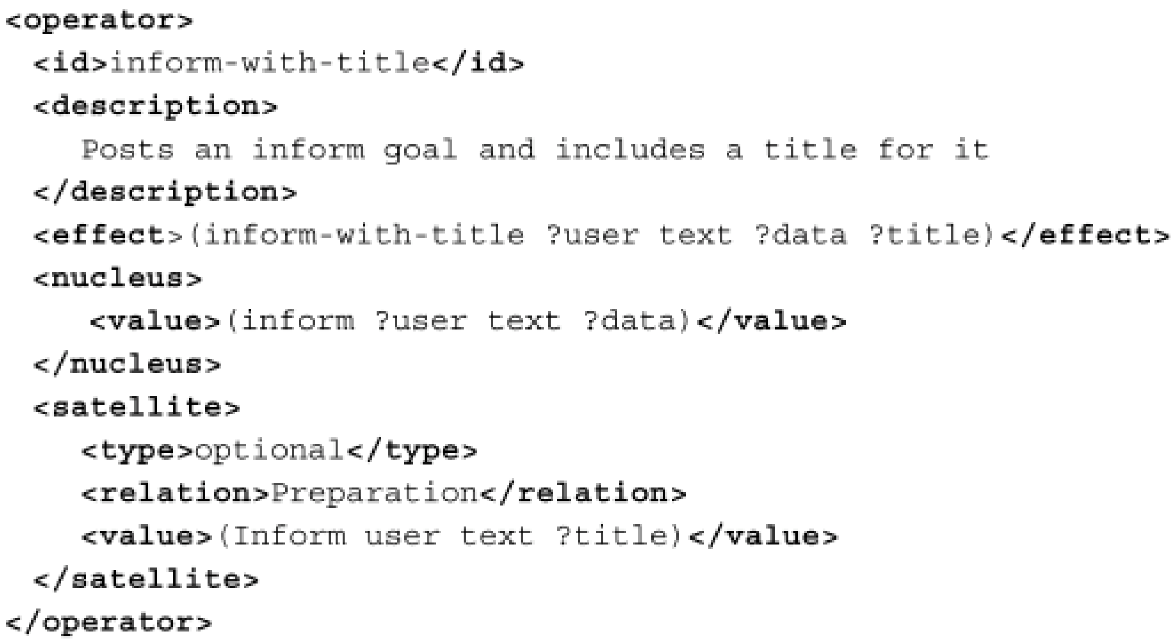


Goal: Provide region report for surveillance operator
Preparation
Context
Provide report for port
title
Parallel
Preparation
Provide report for ship in port
Provide report for piracy-attack in port
title
Weather
Context
Elaboration
Provide All ship Summary
Provide High Risk Ship summary
List all High Risk ship

- The system maintains a library of plans capable of producing such trees top-down from an initial communicative goal.
  1. The plans hierarchically decompose goals
  2. Recursive application of plans until all goals are satisfied and a discourse structure is produced

Example 1: Paris et al. 2010

- Plans access knowledge base and user model and history so that their effects are conditioned by domain knowledge, available data and user preferences.
- Discourse plans decide what content should be included, i.e. they perform content determination too.
  - Focus on coarse-grained units of information.

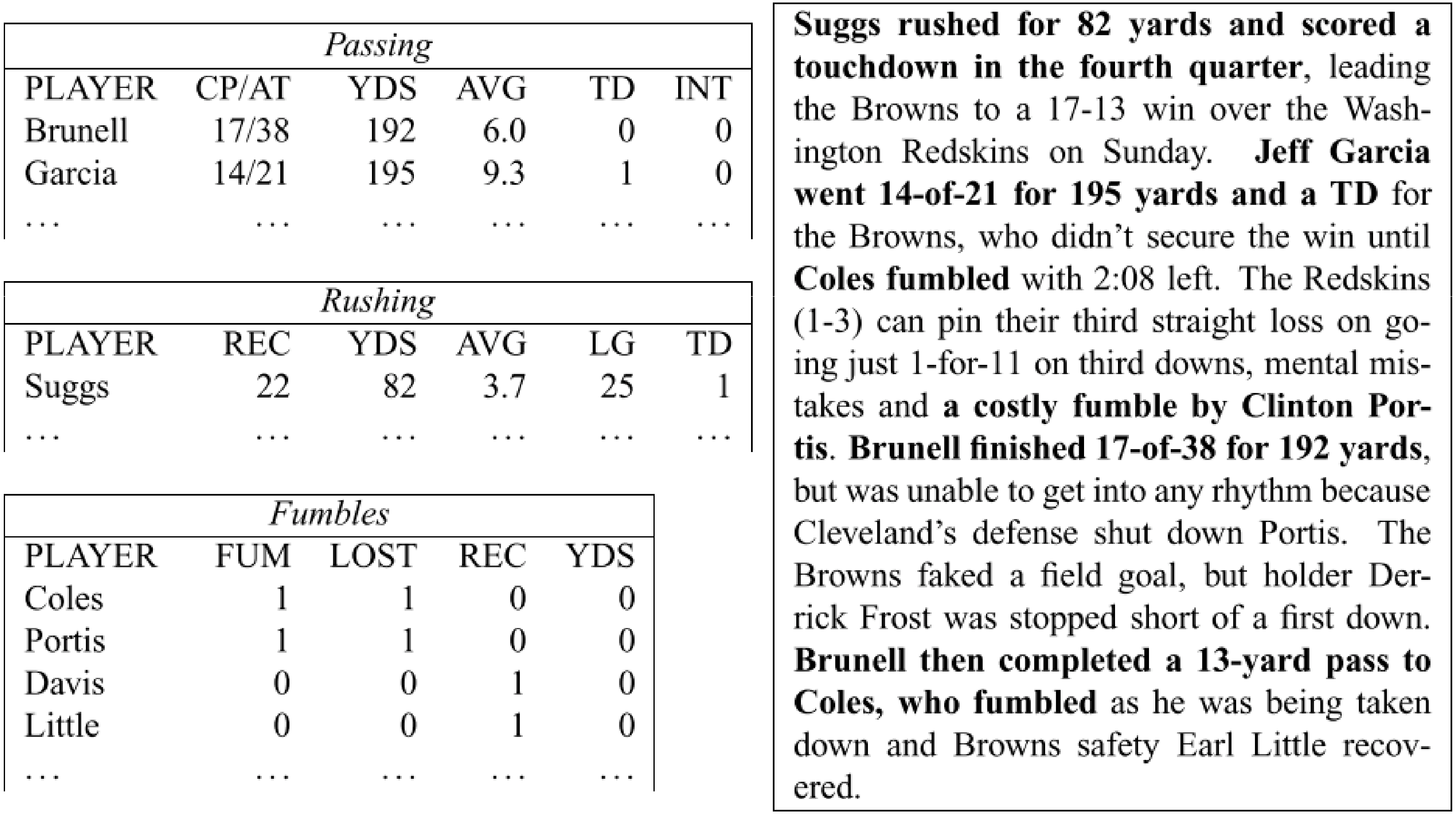


```
<operator>
  <id>inform-with-title</id>
  <description>
     Posts an inform goal and includes a title for it
  </description>
  <effect>(inform-with-title ?user text ?data ?title)</effect>
  <nucleus>
      <value>(inform ?user text ?data)</value>
  </nucleus>
  <satellite>
     <type>optional</type>
     <relation>Preparation</relation>
     <value>(Inform user text ?title)</value>
  </satellite>
</operator>
```

Example 2: Barzilay and Lapata 2005

- Input data:
  - structured -> relational database containing events in football matches. Size: 73,400 rows distributed across 17 tables.
  - Granularity: a row of a table in the DB
- Bottom-up
- Separate task
- Methods: statistics, graph algorithm.

Example 2: Barzilay and Lapata 2005



Passing

PLAYER	CP/AT	YDS	AVG	TD	INT
Brunell	17/38	192	6.0	0	0
Garcia	14/21	195	9.3	1	0
…	…	…	…	…	…

Rushing

PLAYER	REC	YDS	AVG	LG	TD
Suggs	22	82	3.7	25	1
…	…	…	…	…	…

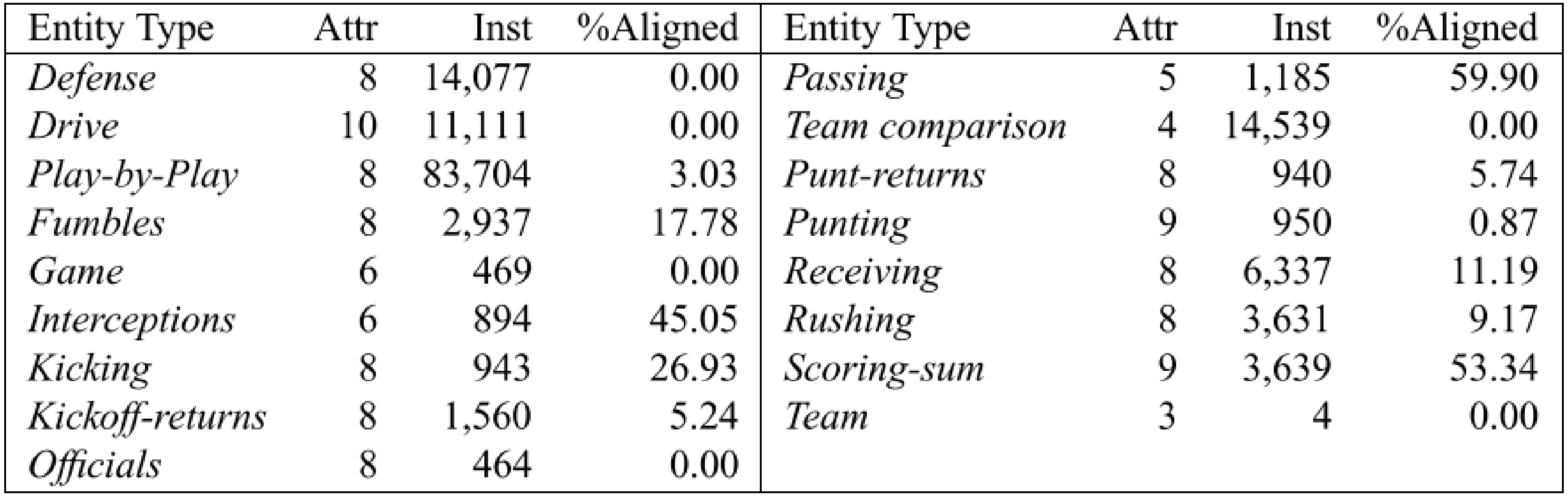
Fumbles

PLAYER	FUM	LOST	REC	YDS
Coles	1	1	0	0
Portis	1	1	0	0
Davis	0	0	1	0
Little	0	0	1	0
…	…	…	…	…

Suggs rushed for 82 yards and scored a touchdown in the fourth quarter, leading the Browns to a 17-13 win over the Washington Redskins on Sunday. **Jeff Garcia went 14-of-21 for 195 yards and a TD** for the Browns, who didn't secure the win until **Coles fumbled** with 2:08 left. The Redskins (1-3) can pin their third straight loss on going just 1-for-11 on third downs, mental mistakes and **a costly fumble by Clinton Portis**. **Brunell finished 17-of-38 for 192 yards**, but was unable to get into any rhythm because Cleveland's defense shut down Portis. The Browns faked a field goal, but holder Derrick Frost was stopped short of a first down. **Brunell then completed a 13-yard pass to Coles, who fumbled** as he was being taken down and Browns safety Earl Little recovered.

Example 2: Barzilay and Lapata 2005

- Database rows are automatically aligned to sentences in a corpus of match reports using simple anchor-based techniques.
  - Overlap between cell values in row and tokens in sentence.
  - This works because proper names and numbers which are easy to align occur with high frequency.



Entity Type	Attr	Inst	%Aligned
Defense	8	14,077	0.00
Drive	10	11,111	0.00
Play-by-Play	8	83,704	3.03
Fumbles	8	2,937	17.78
Game	6	469	0.00
Interceptions	6	894	45.05
Kicking	8	943	26.93
Kickoff-returns	8	1,560	5.24
Officials	8	464	0.00
Passing	5	1,185	59.90
Team comparison	4	14,539	0.00
Punt-returns	8	940	5.74
Punting	9	950	0.87
Receiving	8	6,337	11.19
Rushing	8	3,631	9.17
Scoring-sum	9	3,639	53.34
Team	3	4	0.00

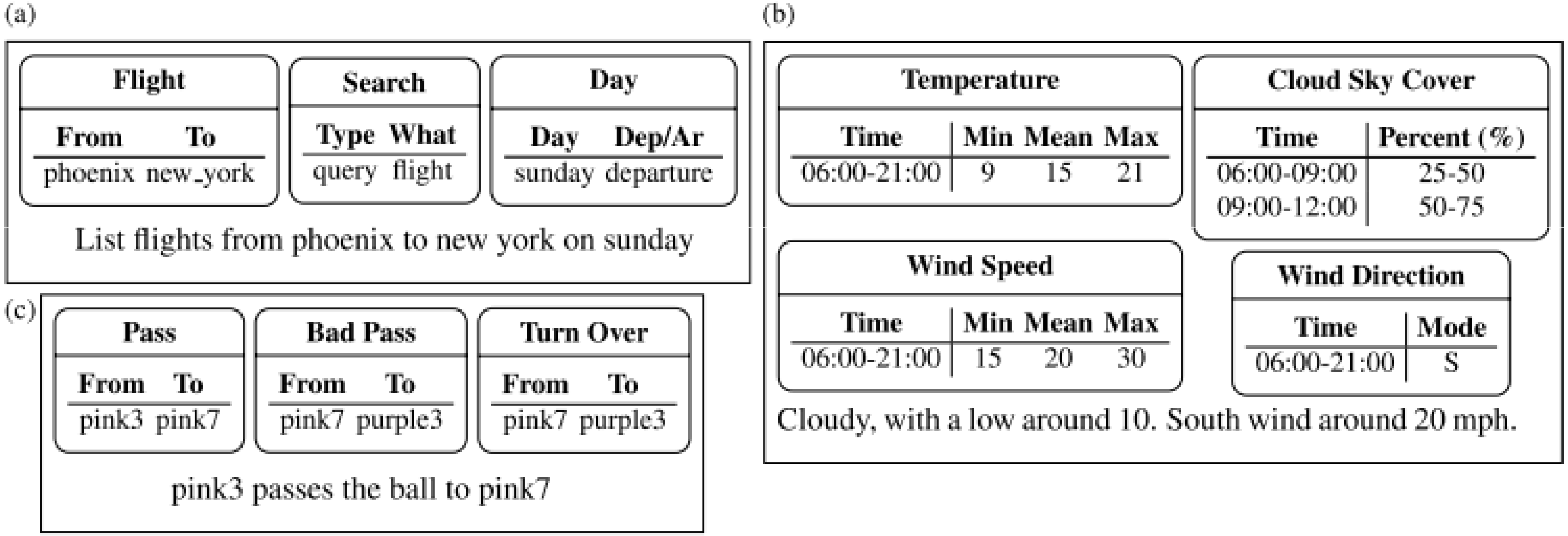
Example 2: Barzilay and Lapata 2005

- Links between a row and a sentence constitute positive examples of selection of the row.
- Collective selection of database rows belonging to a match:
    1. A graph is built where nodes are rows in the database and edges indicate semantic relatedness, i.e. the connected rows share at least one attribute value.
    2. Nodes weights are predictions of a set of models trained using machine learning.
    3. Edge pruning: Discard edges where both rows have different selection distribution across documents.
    4. Edge weighting: use simulated annealing to obtain a global assignment of weights according to node weights and edges.
- Result: weighted constraints between pairs of rows belonging to a match.

Example 3: Konstas and Lapata 2013

- Input data:
  - structured -> relational databases belonging to three domains: Robocup game finals, weather forecasts and air travel.
  - Granularity: a row of a table in the DB
- Data-driven
- Content determination interleaved with ordering and surface realization.
- Methods: statistics, symbolic grammar (PCFG).

Example 3: Konstas and Lapata 2013



(a)

Flight

From	To
phoenix	new_york

Search

Type	What
query	flight

Day

Day	Dep/Ar
sunday	departure

List flights from phoenix to new york on sunday

(b)

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

Cloudy, with a low around 10. South wind around 20 mph.

(c)

Pass

From	To
pink3	pink7

Bad Pass

From	To
pink7	purple3

Turn Over

From	To
pink7	purple3

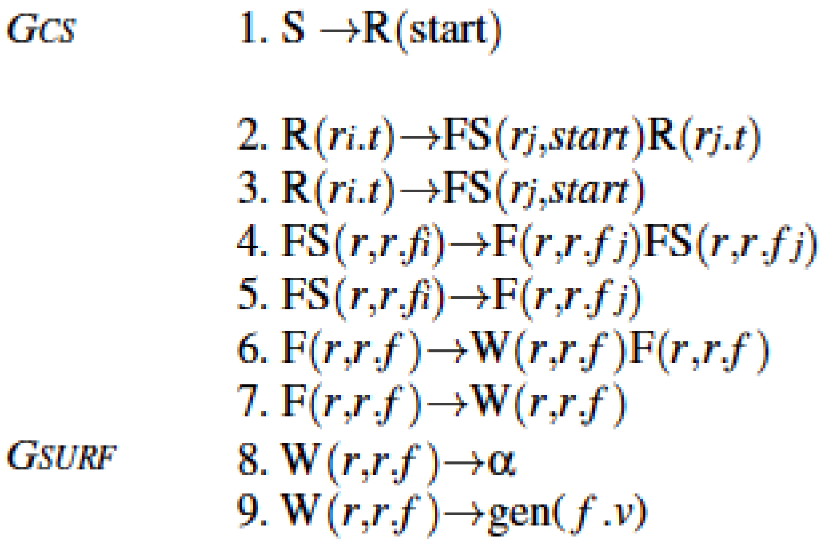
pink3 passes the ball to pink7

Figure 1: Input-output examples for (a) query generation in the air travel domain, (b) weather forecast generation, and (c) sportscasting.

Example 3: Konstas and Lapata 2013

- Probabilistic Context Free Grammar (PCFG) as set of rewrite rules: rewrite structure of DB (rows, cells) into words.

- The weights of the PCFG are estimated by applying the CYK parser and the grammar to texts in a corpus of verbalizations of the DB.
- For each text a hypergraph is built and the weights are updated.
- Generation maximizes the PCFG grammar and an n-gram language model.
- Uses datasets of DB records (manually) paired with texts verbalizing them.



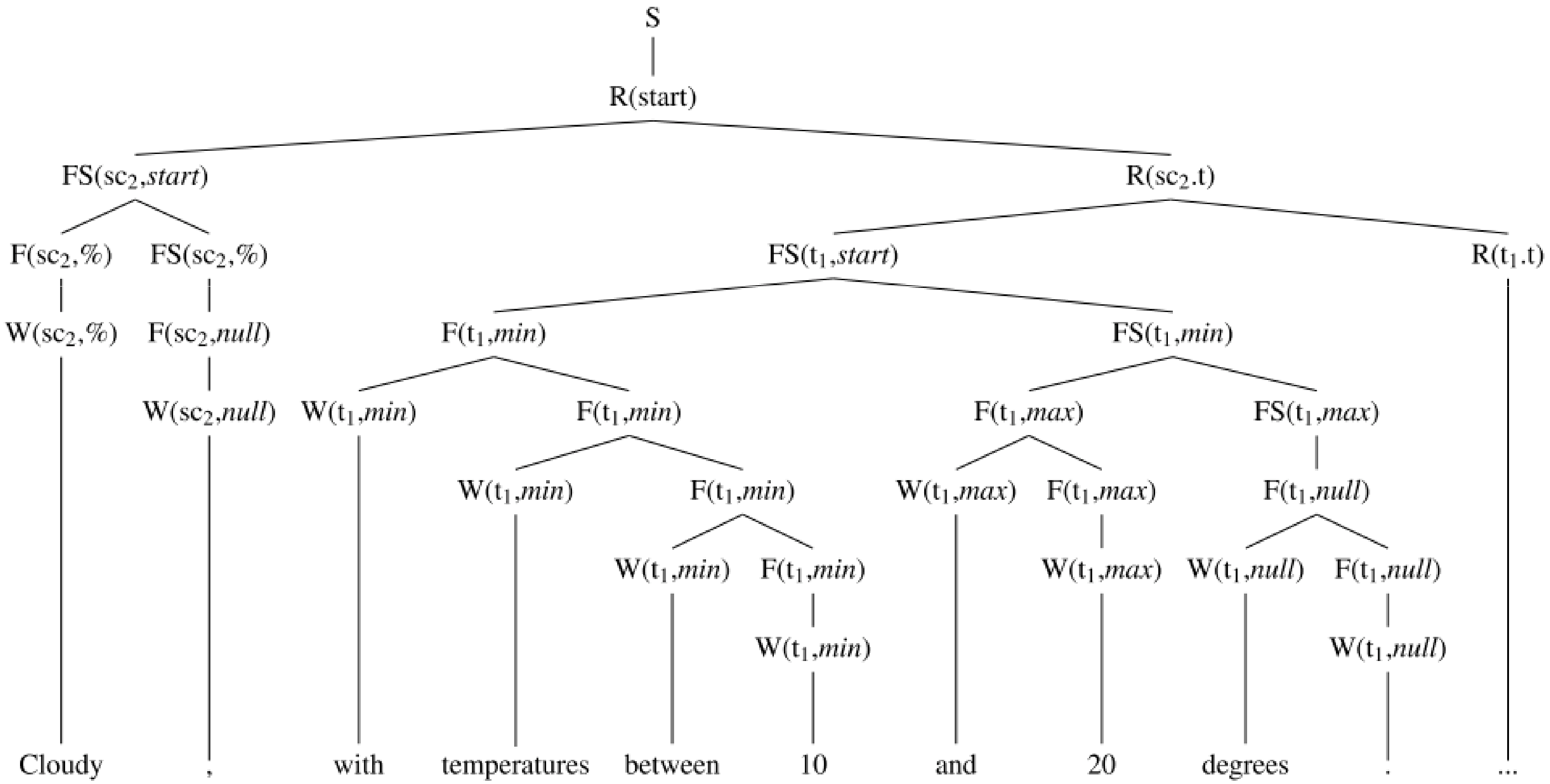
G_{CS}

1. $\mathrm{S} \rightarrow \mathrm{R}(\mathrm{start})$
2. $\mathrm{R}(r_i.t) \rightarrow \mathrm{FS}(r_j, start)\,\mathrm{R}(r_j.t)$
3. $\mathrm{R}(r_i.t) \rightarrow \mathrm{FS}(r_j, start)$
4. $\mathrm{FS}(r, r.f_i) \rightarrow \mathrm{F}(r, r.f_j)\,\mathrm{FS}(r, r.f_j)$
5. $\mathrm{FS}(r, r.f_i) \rightarrow \mathrm{F}(r, r.f_j)$
6. $\mathrm{F}(r, r.f) \rightarrow \mathrm{W}(r, r.f)\,\mathrm{F}(r, r.f)$
7. $\mathrm{F}(r, r.f) \rightarrow \mathrm{W}(r, r.f)$

G_{SURF}

8. $\mathrm{W}(r, r.f) \rightarrow \alpha$
9. $\mathrm{W}(r, r.f) \rightarrow \mathrm{gen}(f.v)$

Example 3: Konstas and Lapata 2013



4. Content Determination from SW Data

Why does it matter?

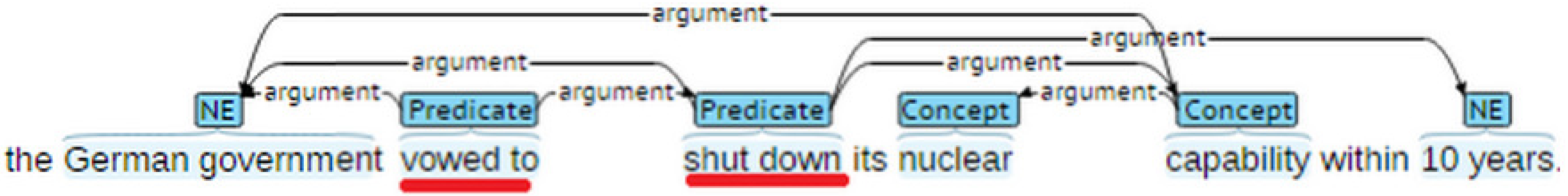
1. A common interface for accessing and reasoning about data:
   - HTTP, URIs, RDF, RDFS, OWL, SPARQL, reasoners, etc.
2. Large (and growing) amounts of Linked Open Data (LOD) belonging to multiple domains.
   - NLG can be used to make data accessible to humans.
3. Explicit semantics facilitate data assessment and document planning.
4. NLG-relevant knowledge can be modeled using SW standards and shared LD publishing standards.
   - E.g. Lemon-encoded lexical resources, NIF corpora.
5. Advances in Information Extraction means that corpora of texts annotated with SW data can be created automatically.
6. More research is needed!

Content Determination from SW data

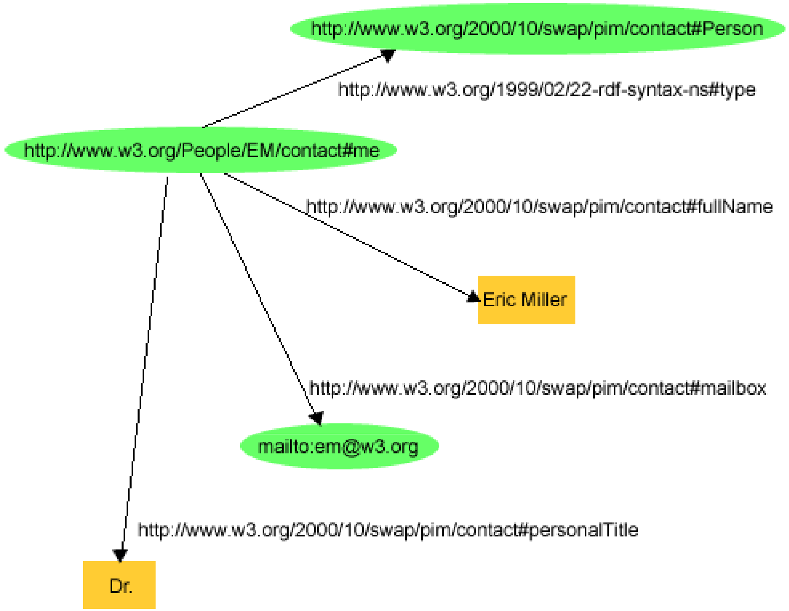
- Four main communicative goals in NLG approaches for SW data:
    i. to say almost all there is to say about some input object (i.e., class, query, constraint, whole graph)
    ii. to verbalize content interactively selected by the user
    iii. To verbalize the most typical facts found in target texts
    iv. To verbalize the most relevant facts according to the context
- No (or trivial) Content Determination in (i) and (ii)
- Templates and schemas used for (iii)
- More elaborate strategies (i.e. statistical methods, graph-based) could be used.

Annotating texts with SW data

- Entity Linking
  - Named entity recognition + disambiguation against a dataset
  - Wikipedia/DBPedia/BabelNet
- Coreference resolution
- Annotation of relations
  - Deep parsers
    - Boxer, Mate tools, Abstract Meaning Representation (AMR) parsers.
  - Semantic Role Labelling
    - FrameNet, VerbNet, PropBank.



Selecting RDF data


http://www.w3.org/2000/10/swap/pim/contact#Person
http://www.w3.org/1999/02/22-rdf-syntax-ns#type
http://www.w3.org/People/EM/contact#me
http://www.w3.org/2000/10/swap/pim/contact#fullName
Eric Miller
http://www.w3.org/2000/10/swap/pim/contact#mailbox
mailto:em@w3.org
http://www.w3.org/2000/10/swap/pim/contact#personalTitle
Dr.

- RDF can be viewed as a graph.
  - Properties as edges
  - Entity-sharing as edges
  - Most LD datasets are huge graphs!
- Granularity of selection:
  - Single triple != statement
  - Blank nodes
  - But N-ary relations are expressed with multiple triples
- Statistical methods:
  - Ontologies provide features for the creation of models based on data, e.g. ontological types.

Conclusions

- Content determination is very important (in many applications, the ability of the machine to find relevant material is perhaps more important than how the material is stated)
- Good content determination often involves specialised domain reasoning
- Statistical approaches tend to learn CD as a part of a simple complete NLG pipeline.
- More and more NLG applications are starting from SW data.

References

- Gabor Angeli, Percy Liang and Dan Klein, “A Simple Domain-Independent Probabilistic Approach to Generation” *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 502–51, 2010.
- Regina Barzilay and Mirella Lapata, “Modeling Local Coherence: An Entity-Based Approach”, *Computational Linguistics* Volume 34, Number 1, 2008.
- Anja Belz, “Automatic Generation of Weather Forecast Texts Using Comprehensive Probabilistic Generation-Space Models”, *Natural Language Engineering*, 14 (4). pp. 431-455, 2008.
- Bontcheva, Kalina, and Yorick Wilks. "Automatic report generation from ontologies: the MIAKT approach." Natural Language Processing and Information Systems. Springer Berlin Heidelberg, 2004. 324-335.
- Bontcheva, Kalina. "Generating tailored textual summaries from ontologies." The Semantic Web: Research and Applications. Springer Berlin Heidelberg, 2005. 531-545.
- Bouayad-Agha, N., Casamayor, G., & Wanner, L. (2011). Content selection from an ontology-based knowledge base for the generation of football summaries (pp. 72–81). ENLG '11 Proceedings of the 13th European Workshop on Natural Language Generation.
- Bouayad-Agha, Nadjet, et al. "From Ontology to NL: Generation of multilingual user-oriented environmental reports." Natural Language Processing and Information Systems. Springer Berlin Heidelberg, 2012. 216-221.
- Bouttaz, Thomas, et al. "A policy-based approach to context dependent natural language generation." Proceedings of the 13th European Workshop on Natural Language Generation. Association for Computational Linguistics, 2011.
- Carenini, Giuseppe, and Johanna D. Moore. "Generating and evaluating evaluative arguments." Artificial Intelligence 170.11 (2006): 925-952.
- Nina Dethlefs and Heriberto Cuayahuitl, “Hierarchical Reinforcement Learning and Hidden Markov Models for Task-Oriented Natural Language Generation” *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*: short papers, pages 654–659, 2011.
- O’Donnell, Mick, et al. "ILEX: an architecture for a dynamic hypertext generation system." Natural Language Engineering 7.03 (2001): 225-250.

- Hovy, Eduard H. "Automated discourse generation using discourse structure relations." Artificial intelligence 63.1 (1993): 341-385.
- Kelly, Colin, Ann Copestake, and Nikiforos Karamanis. "Investigating content selection for language generation using machine learning." Proceedings of the 12th European Workshop on Natural Language Generation. Association for Computational Linguistics, 2009.
- Ioannis Konstas and Mirella Lapata, “A Global Model for Concept-to-Text Generation” *Journal of Artificial Intelligence Research* 48 pp305-346, 2013.
- Marciniak, Tomasz, and Michael Strube. "Beyond the pipeline: Discrete optimization in NLP." Proceedings of the Ninth Conference on Computational Natural Language Learning. Association for Computational Linguistics, 2005.
- Kathleen McKeown, *Text Generation*, Cambridge University Press, 1992.
- Mellish, Chris, and Jeff Z. Pan. "Natural language directed inference from ontologies." Artificial Intelligence 172.10 (2008): 1285-1315.
- Marie Meteer, "Bridging the 'Generation Gap' between Text Planning and Linguistic Realization” *Computational Intelligence.* 7(4) Special issue on Natural Language Generation, 1992.
- Cécile Paris, Nathalie Colineau, Andrew Lampert, and Keith Vander linden. 2010. Discourse planning for information composition and delivery: A reusable platform. Nat. Lang. Eng. 16, 1 (January 2010), 61-98.
- François Portet, Ehud Reiter, Albert Gatt, Jim Hunter, Somayajulu Sripada, Yvonne Freer and Cindy Sykes, “Automatic generation of textual summaries from neonatal intensive care data”. *Proceedings of the 11th Conference on Artificial Intelligence in Medicine*, AIME 2007, pages 227-236, 2007.
- Rambow, Owen. "Domain communication knowledge." Fifth International Workshop on Natural Language Generation. 1990.
- Edhu Reiter and Robert Dale, “Building natural language generation systems” Cambridge University Press, 2000.
- Verena Rieser, Oliver Lemon and Xingkun Yu, “Optimising Information Presentation for Spoken Dialogue Systems”, *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1009–1018, 2010.
- Stuart Shieber, “The Problem of Logical-Form Equivalence”, *Computational Linguistics,* Vol 19, No 1, 1993.
- Yaji Sripada, Ehud Reiter, Jim Hunter and Jin Yu, “Generating English Summaries of Time Series Data Using the Gricean Maxims”, In Proceedings of KDD 2003, pp 187-196, 2003.
- Michael White, “Statistical Natural Language Generation Part I: Content and Sentence Planning”, 2010 http://winterfest.hcsnet.edu.au/files2/2010/winterfest/white-bowral-part1v2.pdf
- R. M. Young, J. D. Moore, “DPOCL: A Principled Approach to Discourse Planning”, in *Proceedings of the 7th International Workshop on Natural Language Generationy*, Kinnebunkport, ME, 13-20, 1994.