

GRAPH: The Costs of Redundancy in Referring Expressions

Emiel Kraahmer

Tilburg University, The Netherlands
e.j.kraahmer@uvt.nl



Mariët Theune

University of Twente, The Netherlands
mtheune@utwente.nl



Jette Viethen

Macquarie University, Australia
jviethen@ics.mq.edu.au



Iris Hendrickx

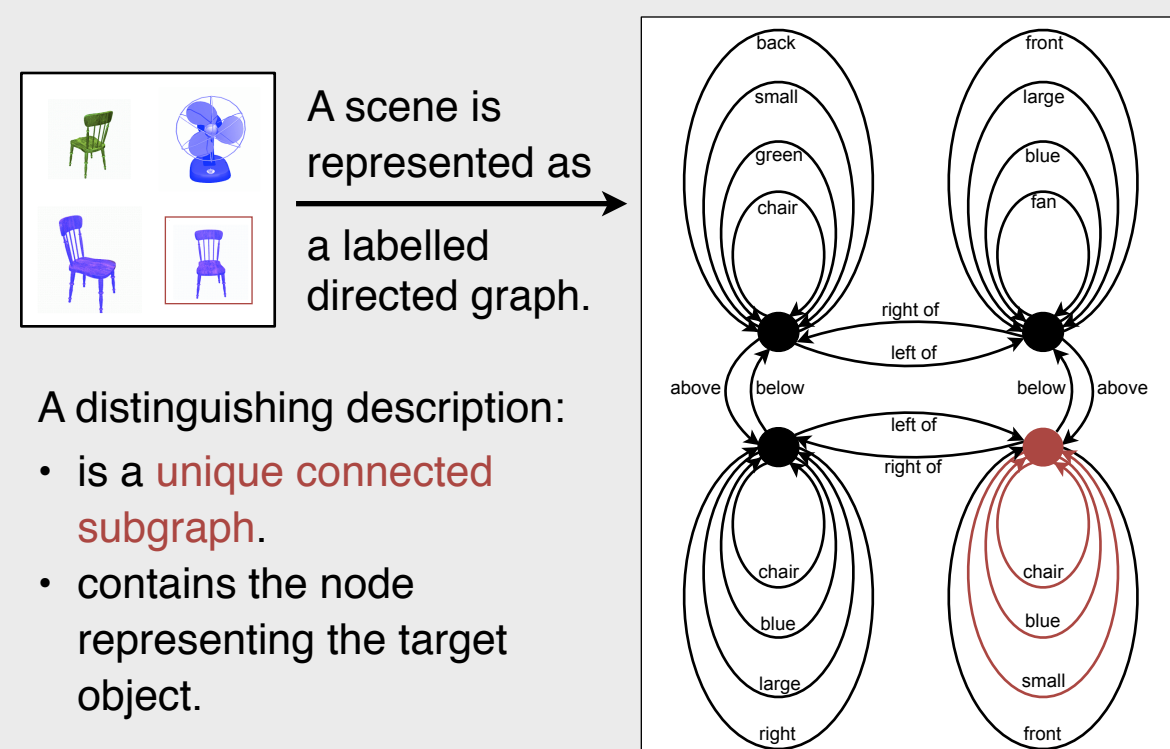
University of Antwerp, Belgium
iris.hendrickx@ua.ac.be



Redundancy in REG

- We present a corpus-based approach to mimicking human use of redundant properties in referring expressions.
- At ASGRE 2007 there was a “trend for the [human-likeness] score [...] to decrease as the proportion of minimal descriptions increases” (Belz and Gatt 2007).
- Existing Referring Expression Generation algorithms (e.g. the Incremental Algorithm (IA)) don’t allow redundancy in a principled way ...
- ... but the Graph-Based Framework (Kraahmer et al. 2003) provides better control over content selection via two fine-grained parameters.

The Graph-Based Algorithm



The algorithm:

- does a depth-first search over the edges.
- uses a cost function over the edges as heuristic.
- returns the cheapest distinguishing description.

Cost Functions

Suppose we have two distinguishing descriptions:

d1 – The front-facing chair
d2 – The small blue one

and three different cost functions over the properties:

cost function	CHAIR	FRONT	SMALL	BLUE
#1	1	12	11	11
#2	1	12	2	3
#3	1	4	2	3

then the algorithm chooses between d1 and d2 as follows:

cost function #1 → cost(d1) = 13, cost(d2) = 22
 cost function #2 → cost(d1) = 13, cost(d2) = 5
 cost function #3 → cost(d1) = cost(d2) = 5

If two referring expressions have the same cheapest cost (#3), the algorithm chooses the one it encounters first.

Property Orderings

Which description is found first is determined by the order in which properties are considered for inclusion.

So, if the cost function doesn’t arbitrate between the two descriptions, the property ordering becomes crucial:

Ordering 1: [CHAIR, SMALL, FRONT, BLUE]

→ d1 – The front-facing chair is chosen.

Ordering 2: [SMALL, CHAIR, BLUE, FRONT]

→ d2 – The small blue one is chosen.

Redundancy in the Graph-Based Algorithm

The cost function has to be monotonically increasing, but NOT strictly. So,

d3 – The blue front-facing chair

can never be cheaper than

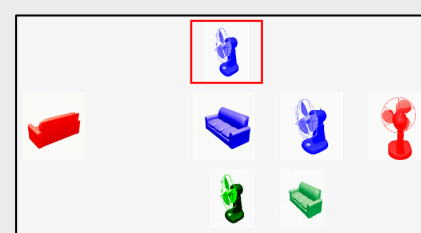
d1 – The front-facing chair.

d3, containing the redundant BLUE, will be returned only if:

- cost(BLUE) = 0 (i.e. d1 and d3 have the same cost); and
- the property ordering favours BLUE (i.e. d3 is found first).

The TUNA Data

- The largest data set of human-produced distinguishing descriptions.
- Two domains: *Furniture* and *People*.
- Used for the ASGRE 2007 and REG 2008 competitions.
- A property count in the combined training and development sets revealed that very frequent properties were often used redundantly.



Tuning the Algorithm

We tried the following parameter settings:

Cost functions:

- Simple Costs:** All properties cost 1 (baseline).
- Stochastic Costs:** determined by property frequency.
- Free–Stochastic:** The most frequent properties are free, the rest have stochastic cost.
- Free–Naïve:** The most frequent properties are free, the least frequent cost 2, and the rest cost 1.

Property Orderings:

- Random** (baseline)
- Cost-based:** properties are tried in stochastic order from cheapest to most expensive.

This results in 8 combinations to be tested.

Our Hypothesis:

The combination 4+B should outperform the other settings on human-likeness, because it allows properties occurring very frequently in the corpus to be used redundantly.

Attribute Selection Evaluation

The purpose of the graph-based algorithm is semantic content selection for referring expressions rather than full surface realisation.

Measures

DICE: coefficient of similarity between a candidate and a reference set of attributes.

$$DICE(A, B) = \frac{2 \times |A \cap B|}{|A| + |B|}$$

MASI: measure of set similarity with a monotonicity coefficient δ favouring subsets.

$$MASI(A, B) = \delta \times \frac{|A \cap B|}{|A \cup B|}$$

A–A (attribute accuracy): ratio of descriptions matching the corpus exactly in content.

MIN (minimality): ratio of descriptions containing no redundant properties.

Results for the Development Sets

	Furniture				People				Combined			
	DICE	MASI	A–A	MIN	DICE	MASI	A–A	MIN	DICE	MASI	A–A	MIN
1+A	.61	.32	.12	.29	.59	.36	.24	.00	.60	.34	.18	.16
1+B	.61	.31	.12	.29	.66	.42	.24	.00	.63	.36	.18	.16
2+A	.71	.47	.31	.11	.66	.42	.24	.00	.69	.45	.28	.06
2+B	.69	.44	.28	.16	.66	.42	.24	.00	.68	.43	.26	.09
3+A	.80	.58	.45	.00	.68	.41	.19	.00	.74	.51	.33	.00
3+B	.80	.58	.45	.00	.72	.48	.28	.00	.76	.54	.37	.00
4+A	.80	.59	.48	.00	.59	.34	.18	.00	.70	.48	.34	.00
4+B	.80	.59	.48	.00	.72	.48	.28	.00	.76	.54	.39	.00

Based on these results, we submitted the graph-based algorithm with the parameter setting 4+B as **GRAPH 4+B** to the TUNA-AS task at REG 2008.

Evaluation of End-to-end REG

To be able to submit GRAPH 4+B to the end-to-end TUNA-REG task at REG 2008, we used a template-based surface realiser provided by the competition organisers.

Measures

EDIT: Levenshtein string-edit distance between the generated word string and the human reference description.

S–A (string accuracy): ratio of descriptions matching the corpus exactly at string level.

Results for the Development Sets

GRAPH	Furniture		People		Combined	
	EDIT	S–A	EDIT	S–A	EDIT	S–A
1+A	5.90	.04	6.54	.00	6.20	.02
1+B	5.89	.04	6.78	.00	6.30	.02
2+A	5.06	.05	6.78	.00	5.85	.03
2+B	5.19	.05	6.78	.00	5.92	.03
3+A	4.90	.05	6.79	.00	5.77	.03
3+B	4.90	.05	6.96	.00	5.84	.03
4+A	4.61	.05	6.56	.00	5.51	.03
4+B	4.61	.05	6.96	.00	5.69	.03

Ranks at REG 2008

Attribute Selection (AS)

14 distinct systems were submitted to the TUNA-AS task at REG 2008. Our GRAPH 4+B system scored as follows:

	DICE	MASI	A–A	MIN
Furniture	1st (shared with 1)	1st (shared with 1)	2nd	
People	3rd	3rd	2nd	
Combined	2nd	2nd	2nd	2nd = last shared with 9

End-to-end REG

The official evaluation results for the GRAPH 4+B system combined with the provided realiser:

	intrinsic measures (out of 14)				extrinsic measures (out of 10)		
	EDIT	S–A	BLEU	NIST	reading time	identification time	error rate
Furniture	1st	1st			6th	7th	3rd
People	13th	last (shared with 7)			4th	8th	3rd (shared with 1)
Combined	5th	2nd	3rd	5th	5th	9th	2nd (shared with 1)

Note that only high differences in ranks were statistically significant. For a full report of the results of the REG 2008 TUNA tasks see Gatt et al. (2008).

Discussion and Conclusions

Attribute Selection

- With *Free–Naïve* costs and *Cost-based* property ordering (GRAPH 4+B) the algorithm achieves high ranks on the human-likeness measures at REG 2008.
- Allowing redundancy boosts performance:
 - Functions with free properties outperform others.
 - Property ordering *B* mostly outperforms baseline *A*.
 - Minimality is negatively correlated to human-likeness.
- Varying costs outperform static *Simple Costs*.
- In combination with cost function *B*, *Free–Naïve* scores equal with the more principled *Free–Stochastic*.

End-to-end REG

- The REG 2008 ranks are inconclusive.
- Combining human-like content selection with a simple off-the-self realiser is not good enough:
 - S–A* is dramatically lower than *A–A*.
 - In the *People* domain, *EDIT* scores gets worse as *DICE*, *MASI* and *A–A* scores get better.
- This is most likely due to:
 - the low chance of an exact string match compared to an exact content match.
 - idiosyncrasies of the realiser not matching human realisation patterns.

References

- Belz, A. and Gatt, A. (2007). The attribute selection for GRE challenge: Overview and evaluation results. In *Proceedings of UCNLG+MT*, 75–83.
- Gatt, A., Belz, A. and Kow, E. (2008). The TUNA Challenge 2008: Overview and evaluation results. In *Proceedings of INLG*, Columbus, OH.
- Kraahmer, E., van Erk, S. and Verleg, A. (2003). Graph-based generation of referring expressions. In *Computational Linguistics 29(1)*, 53–72.
- Viethen, J., Dale, R., Kraahmer, E., Theune, M. and Touset, P. (2008). Controlling redundancy in referring expressions. In *Proceedings of LREC*, Marrakech, Morocco.