On The Feasibility of Open Domain Referring Expression Generation Using Large Scale Folksonomies

Fabián Pacheco, **Pablo Duboue** and Martín Domínguez

Facultad de Matemática, Astronomía y Física
Universidad Nacional de Córdoba
Córdoba, Argentina

Referring Expression Generation (REG)

- Classic NLG problem
  - **Input:** set of entities (with a distinguished element), set of triples pertaining to the entities.
  - **Output:** a Definite Description, i.e., a set of *positive triples* and *negative triples*.
  - Focus (among other things) on running time **efficiency**.
- Question: does efficiency matters nowadays?
  - Yes, it does.
  - We used a large scale *folksonomy* (DBpedia) and a set of naturally occurring entities (from Wikinews).

Can REG Help Summarization?

- Do we have data for the relevant entities?
  - Yes, roughly 50% of the time.
  - We used anaphora training data and looked it up on DBpedia by hand.
- Do we have **discriminant** data for relevant entities?
  - Yes, roughly 80% of the time.
  - Measured on Wikinews, Cohen's κ of 79% (small evaluation size, though).
- Are classic REG algorithms enough?
  - *Maybe not,* they either fail to produce an output or return a poor description in 60%+ of the cases.
  - But there is hope and our evaluation needs to be extended.

About The Authors

Possible Application To Multi-document Summarization

Use REG to fix anaphoric references drafted from different documents (similar to [Siddharthan et al., 2011])

- Excerpt from Columbia Newsblaster:

Thousands of cheering, flag-waving Palestinians gave Palestinian Authority President Mahmoud Abbas an enthusiastic welcome in Ramallah on Sunday, as he told them triumphantly that a "Palestinian spring" had been born following his speech to the United Nations last week. The ***president*** *pressed Israel, in unusually frank terms, to reach a final peace agreement with the Palestinians, citing the boundaries in place on the eve of the June 1967 Arab-Israeli War as the starting point for negotiation about borders.*

Three Single Referent REG Algorithms

- DR [Dale and Reiter, 1995]
  - A classic algorithm.
  - Greedy approach, use a **default ordering**.
- Gardent [Gardent, 2002]
  - An algorithm generating negations.
  - Constraint satisfaction programming.
- Full Brevity (FB) [Bohnet, 2007]
  - More exhaustive search of the solution space

Data: DBpedia

- DBpedia [Bizer et al., 2009] is an ontology curated from Wikipedia infoboxes
  - Infoboxes are the small tables containing structured information at the top of most Wikipedia pages.
  - We used “Ontology Infobox Properties” which contains 1,7520,158 triples (for English).
  - *We missed Ontology Infobox **Types***.

Experiments With Anaphora Resolution Training Data

- Hand-annotated corpus [Hasler et al., 2006]
  - 74 documents, 239 coreference chains.
  - 44% in DBpedia
  - 16 documents usable for REG eval (40 REG tasks).
- Failure rate
  - DR: 12 (30%), Gardent: none (0%), FB: 23 (57.5%).
    - Lack of unique differentiating triples.
    - FB ran out of memory multiple times.
- Execution timings
  - DR and Gardent, comparable; FB 16x slower.
- Discard FB

Experiments With Wikinews-derived REG Tasks

- Wikinews, a news service operated as a wiki
  - News articles interspersed with *interwiki* links.
    - Entities disambiguated.

```
Former [[New Mexico]] {{w|Governor of New
Mexico|governor}} {{w|Gary Johnson}} ended his
campaign for the {{w|Republican Party (United
States)|Republican Party}}
```

- Finding people and organizations
  - Entity has “birth date”? $\Rightarrow$ person
  - Entity has “creation date”? $\Rightarrow$ organization.
  - 4,230 tasks (17,814 runs) for people and 12,998 (44,080) for organizations.

Wikinews Timings And Failure Rates

- Failure Rates
  - People
    - DR: 2.8%, Gardent 2% (negations on 14%).
  - Organizations
    - DR: 30.8%, Gardent 0% (negations on 12%).
- Execution Timings
  - For people, Gardent was 46x slower.
  - For organizations, Gardent was 29x slower.
  - DR took 3’ for the 44,080 runs for organizations.

Wikinews Human Evaluation

- Evaluating referring expressions is hard.
  - Open Domain: the judges need to be acquainted with all entities in the training set.
- Inter-annotator agreement
  - Random sample of 20 runs, two annotators.
  - Cohen's κ of 60% for annotating DD results.
  - κ of 79% for determining whether the folksonomy had enough information to build a satisfactory DD.
- Final evaluation
  - Extended to 60 runs (one annotator).
  - DR: 41.6% accuracy; Gardent: 43.4% accuracy.
  - Folksonomy contained enough information: 81.6%.

Issues

- DR algorithm issues
  - Default ordering strategy not stable across different subtypes (e.g., politicians vs. musicians).
  - Recent paper might help (Koolen et al. at INLG'12).
- Gardent's algorithm issues
  - Sometimes it selects a bad triple (an obscure fact).
  - A negative piece of information could just be a missing piece of information.
  - Example: **China** vs. $\{$ Peru and Taiwan $\}$
    - "the place where they do not speak Chinese"
- Robust NLG for noisy (ontological) inputs.

Conclusions

- A folksonomy can enable traditional NLG referring expression generation for Open Domain tasks.
- Three tasks remain:
  - Dealing with missing information.
    - *smart default values*, ontological siblings.
  - Estimating salience for ontological information.
    - Search engine salience.
  - Transform the extracted triples into actual text
    - Custom-made grammar.

Backup Slides

Efforts to automate this task in NLG [Gatt et al., 2007] have taken an approach similar to machine translation BLEU scores [Papinini et al., 2001], for example, by asking multiple judges to produce referring expressions for a given scenario. These settings usually involve images of physical objects and relate to small ontologies. While such an approach could be adapted to the

Intro

- What is Referring Expression Generation (REG)
  - Input: (generation from **data**), ontological information about the referents
  - Output: Definite Descriptions (DD), set of *positive triples* and a set of *negative triples*,
  - Lot of attention in NLG
    - early work: using custom-tailored ontologies
    - recent years: [Belz et al., 2010] "Open Domain Referring Expression Generation," (OD REG), properties come from a *folksonomy*, a large-scale volunteer-built ontology.
- Two sets of experiments:
  - one with anaphora resolution training information

– roughly half of the entities annotated in the documents were present in the folksonomy
– sets of distractors from Wikinews
– 40k referring expression tasks.

References

[Belz et al., 2010] Belz, A., Kow, E., Viethen, J., and Gatt, A. (2010). Generating referring expressions in context: The grec task evaluation challenges. In Krahmer, E. and Theune, M., editors, *Empirical Methods in Natural Language Generation*, volume 5790 of *Lecture Notes in Computer Science*, pages 294–327. Springer.

[Bizer et al., 2009] Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S. (2009). DBpedia-a crystallization point for the web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):154–165.

[Bohnet, 2007] Bohnet, B. (2007). is-fbn, is-fbs, is-iac: The adapexpressions in order to produce expressions like humans do. *MT Summit XI, UCNLG+ MT*, pages 84–86.

[Dale and Reiter, 1995] Dale, R. and Reiter, E. (1995). Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263.

[Gardent, 2002] Gardent, C. (2002). Generating minimal definite descriptions. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 96–103. Association for Computational Linguistics.

[Gatt et al., 2007] Gatt, A., Sluis, I. V. D., and Deemter, K. V. (2007). Evaluating algorithms for the generation of referring expressions using a balanced corpus. In *Proceedings of the Eleventh European Workshop on Natural Language Generation*, pages 49–56. Association for Computational Linguistics.

[Hasler et al., 2006] Hasler, L., Orasan, C., and Naumann, K. (2006). NPs for events: Experiments in coreference annotation. In *Proceedings of the 5th edition of the International Conference on Language Resources and Evaluation (LREC2006)*, pages 1167–1172.

[Papinini et al., 2001] Papinini, K., Roukos, S., Ward, T., and Zhu, W.-J. (2001). Bleu: a method for automatic evaluation of machine translation. Technical report, IBM.

[Siddharthan et al., 2011] Siddharthan, A., Nenkova, A., and McKeown, K. (2011). Information status distinctions and referring expressions: An empirical study of references to people in news summaries. *Computational Linguistics*, 37(4):811–842.