On the Robustness of

Standalone Referring Expression Generation

Algorithms Using RDF Data

Pablo Duboue, Martín Domínguez and Paula Estrella

Facultad de Matemática, Astronomía y Física

Universidad Nacional de Córdoba

Córdoba, Argentina

Our Work at a Glance

- The need of robust NLG algorithms.

- 3 algorithms of REG on anachronistic input.
  - On old data, produce a ref. expression, check whether holds.
  - We do not analyze whether the referring expression is good.
  - We have evaluated their goodness in [Pacheco et al., 2012].

- We found poor results with marginal differences among the algorithms.
  - Gardent's algorithm might be ahead but using closed world assumptions.
  - Nice task and problem, worth extending.

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 Descriptions (DD), set of *positive triples* and a set of *negative triples*.

- Question: how good are referring expressions over erroneous data?
  - Naturally occurring errors using anachronistic data.

- Example: distinguish *Paul McCartney* from *Ringo Starr*, *John Lennon*, *George Harrison*
  - NOT associated musical artist: Plastic Ono Band

Three REG Algorithms

- DR [Dale and Reiter, 1995]

  – Greedy approach, use a **default ordering**.

- Gardent [Gardent, 2002]

  – Constraint satisfaction programming

    ∗ Also needs a default ordering.

  – Generates negations.

- Graph [Krahmer et al., 2003]

  – Graph isomorphism.

    ∗ Also needs a default ordering.

Experiments With Wikinews-derived REG Tasks

- **Wikinews, news articles with *interwiki* links.**

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

- **Focus on people and organizations**

Algorithm	Execution Errors	Dice	Omission Errors	Inclusion Errors
People – Entity has "birth date"? ⇒ person (3,051 tasks)				
Incremental	232 (5%)	0.48	1,406 (50%)	145 (5%)
Gardent	0 (0%)	0.58	1,089 (36%)	554 (18%)
Graph	15 (0%)	0.38	1,870 (62%)	20 (0%)
Organizations – Entity has "creation date"? ⇒ organization (2,370 tasks)				
Incremental	1,386 (45%)	0.69	305 (31%)	3 (0%)
Gardent	829 (27%)	0.70	338 (22%)	357 (23%)
Graph	934 (31%)	0.06	1,347 (94%)	2 (0%)

Implementation Details

- Alusivo: an Open Source implementation of REG algorithms
  - `https://github.com/DrDub/Alusivo`
  - Java, Maven, RDF-based

- Interface

  - `public ReferringExpression resolve(URI referent, List<URI> confusors, RepositoryConnection repo)`

- Libraries
  - Sesame (RDF)
  - ChocoSolver (CSP)
  - jgrapht (Graph algorithms)

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.
  - The mappings between the infoboxes labels to the ontology is done in a wiki itself: `http://mappings.dbpedia.org/`.
  - The source code of the scrapping scripts is also available with all its development history.

- Not to be confused with a new project targeting to provide structured information to Wikipedia, wikidata.

Two Versions: Compared

Type files analysis

Property	3.6	2014
Number of triples	6,173,940	28,031,852
Unique subjects (entities)	1,668,503	4,218,628
Unique objects (types)	250	547
Max objects per subject	6	16

Mapping files analysis

Property	3.6	2014
Number of verbs	1,100	1,370
Number of triples	13,795,664	33,449,633

- However, many entities lost their types
  - From 20,693 Politicians in 3.6, 4,542 are gone (20%-25%).
  - However, the total Politicians in 2014 is 40,343.

Error Prone Task

- ## NAACL 2012 bug

  – DBpedia distributed in two files, we used only one.

- ## MICAI 2015 bug

```
FileInputStream inNew = new FileInputStream("data/people.tuples.new");
Model mNew = Rio.parse(inNew, "http://localhost/", RDFFormat.NTRIPLES);
Repository repNew = new SailRepository(new MemoryStore());
ValueFactory fNew = repNew.getValueFactory();
RepositoryConnection connNew = repNew.getConnection();
connNew.add(mNew);

FileInputStream inOld = new FileInputStream("data/people.tuples.old");
Model mOld = Rio.parse(inOld, "http://localhost/", RDFFormat.NTRIPLES);
Repository repOld = new SailRepository(new MemoryStore());
ValueFactory fOld = repOld.getValueFactory();
RepositoryConnection connOld = repNew.getConnection();
connOld.add(mOld);
```

Learning Orderings

- ## Iberamia, to appear:

  - *Using Robustness to Learn to Order Semantic Properties in Referring Expression Generation,* Duboue and Domínguez.

- ## Intuitions

  - A good ref. expression should refer to stable properties.

- ## Results

  - Robustness helps to learning orderings.
  - But popularity on a folksonomy is a stronger signal.

Conclusions

- DBpedia/Wikinews is a suitable source for doing research on robust REG algorithms.

- Where to go from Here:

  - Better understanding of the parameterization of the algorithms.
  - Distinguish data changes from missing data.
  - More REG algorithms.
  - Better testing of our current implementations.

Backup Slides

References

[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.

[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.

[Krahmer et al., 2003] Krahmer, E., Erk, S., and Verleg, A. (2003). Graph-based generation of referring expressions. *Computational Linguistics*, 29:53–72.

[Pacheco et al., 2012] Pacheco, F., Duboue, P. A., and Domínguez, M. A. (2012). On the feasibility of open domain referring expression generation using large scale folksonomies. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 641–645, Stroudsburg, PA, USA. Association for Computational Linguistics.