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

Fabián Pacheco **Pablo Ariel Duboue**[*] **Martín Ariel Domínguez**
Facultad de Matemática, Astronomía y Física
Universidad Nacional de Córdoba
Córdoba, Argentina

Abstract

Generating referring expressions has received considerable attention in Natural Language Generation. In recent years we start seeing deployments of referring expression generators moving away from limited domains with custom-made ontologies. In this work, we explore the feasibility of using large scale noisy ontologies (folksonomies) for open domain referring expression generation, an important task for summarization by re-generation. Our experiments on a fully annotated anaphora resolution training set and a larger, volunteer-submitted news corpus show that existing algorithms are efficient enough to deal with large scale ontologies but need to be extended to deal with undefined values and some measure for information salience.

1 Introduction

Given an entity[1] (the **referent**) and a set of competing entities (the **set of distractors**), the task of referring expression generation (REG) involves creating a mention to the referent so that, in the eyes of the reader, it is clearly distinguishable from any other entity in the set of distractors. In a traditional generation pipeline, referring expression generation happens at the sentence planning level. As a result, its output is not a textual nugget but a description employed later on by the surface realizer. In this paper, we consider the output of the REG system to be Definite Descriptions (DD) consisting of a set of *positive triples* and a set of *negative triples*, enumerating referent-related properties.

Since the seminal work by Dale and Reiter (1995), REG has received a lot of attention in the Natural Language Generation (NLG) community. However, most of the early work on REG has been on traditional NLG systems, using custom-tailored ontologies. In recent years (Belz et al., 2010) there has been a shift towards what we term "Open Domain Referring Expression Generation," (OD REG), that is, a REG task where the properties come from a *folksonomy*, a large-scale volunteer-built ontology.

In particular, we are interested in changing anaphoric references for entities appearing in sentences drafted from different documents, as done in multi-document summarization (Advaith et al., 2011). For example, consider the following summary excerpt[2] as produced by Newsblaster (McKeown et al., 2002):

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.[3] *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 ne-*

[*] To whom correspondence should be addressed. Email: `pablo.duboue@gmail.com`.

[1] Or set of entities, but not in this work.

[2] From `http://newsblaster.cs.columbia.edu/archives/2011-10-07-04-51-35/web/summaries/ 2011-10-07-04-51-35-011.html`.

[3] *After his stint at UN, Abbas is politically stronger than ever* (haaretz.com, 10/07/2011, 763 words).

2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 641–645,
Montréal, Canada, June 3-8, 2012.

gotiation about borders.[4]

Here the second sentence refers to U.S. president Barack Obama and a referring expression of the form “U.S. president” should have been used. Such expressions depend on the set of distractors present in the text, a requirement that highlights the dynamic nature of the problem. Our experiments extracted thousands of complex cases (such as distinguishing one musician from a set of five) which we used to test existing algorithms against a folksonomy, dbPedia[5] (Bizer et al., 2009). This folksonomy contains 1.7M triples (for its English version) and has been curated from Wikipedia.[6]

We performed two experiments: first we employed sets of distractors derived from a set of documents annotated with anaphora resolution information (Hasler et al., 2006). We found that roughly half of the entities annotated in the documents were present in the folksonomy, which speaks of the feasibility of using a folksonomy for OD REG, given the fact that Wikipedia has strict notability requirements for adding information. In the second experiment, we obtained sets of distractors from Wikinews,[7] a service where volunteers submit news articles interspersed with Wikipedia links. We leveraged said links to assemble 40k referring expression tasks.

For algorithms, we employed Dale and Reiter (1995), Gardent (2002) and Full Brevity (FB) (Bohnet, 2007). Our results show that the first two algorithms produce results in a majority of the referring expression tasks, with the Dale and Reiter algorithm being the most efficient and resilient of the three. The results, however, are of mixed quality and more research is needed to overcome two problems we have identified in our experiments: dealing with undefined information in the folksonomy and the need to incorporate a rough user model in the form of information salience.

In the next section we briefly summarize the three algorithms we employed in our experiments. In Section 3, we describe the data employed. Section 4 contains the results of our experiments and subsequent analysis. We conclude discussing future work.

2 Referring Expression Generation (REG)

REG literature is vast and spans decades of work. We picked three algorithms with the following desiderata: all the algorithms can deal with single entity referents (a significant amount of recent work went into multi-entity referents) and we wanted to showcase a classic algorithm (Dale and Reiter’s), an algorithm generating negations (Gardent’s) and an algorithm with a more exhaustive search of the solutions space (Full Brevity). We very briefly describe each of the algorithms in turn, where R is the referent, C is the set of distractors and P is a list of properties, triples in the form (entity, property, value), describing R:

Dale and Reiter (1995). They assume the properties in P are ordered according to an established criteria. Then the algorithm iterates over P, adding each triple one at a time and removing from C all entities ruled out by the new triple. Triples that do not eliminate any new entities from C are ignored. The algorithm terminates when C is empty.

Gardent (2002). The algorithm uses Constraint Satisfaction Programming to solve two basic constraints: find a set of positive properties P^+ and negative properties P^-, such that all properties in P^+ are true for the referent and all in P^- are false, and it is the smaller $P^+ \cup P^-$ such that for every $c \in C$ there exist a property in P^+ that does not hold for c or a property in P^- that holds for c.[8]

Full Brevity (Bohnet, 2007). Starting from a state E of the form (L, C, P) with $L = \emptyset$ (selected properties), it keeps these states into a queue, where it loops until $C = \emptyset$. In each loop it generates new states (added to the end of the queue), as follows: given a state $E = (L, C, P)$ for each $p \in P$, if p removes elements rem from C, it adds $(L \cup \{p\}, C - rem, P - \{p\})$, otherwise $(L, C, P - \{p\})$.

3 Data

dbPedia. dbPedia (Bizer et al., 2009) is an ontology curated from Wikipedia infoboxes, small tables containing structured information at the top of most Wikipedia pages. The version employed in this paper (“Ontology Infobox Properties”) contains 1,7520,158 triples. Each

[4] *Obama prods Mideast allies to embrace reform, make peace* (Washington Post, 10/07/2011, 371 words).

[5] http://dbpedia.org

[6] http://wikipedia.org

[7] http://wikinews.org

[8] We employed the Choco CSP solver Java library: http://www.emn.fr/z-info/choco-solver/.

```
Former [[New Mexico]] {{w|Governor of New
Mexico|governor}} {{w|Gary Johnson}} ended
his campaign for the {{w|Republican Party
(United States)|Republican Party}} (GOP)
presidential nomination to seek the backing
of the {{w|Libertarian Party (United
States)|Libertarian Party}} (LP).
```

Figure 1: Wikinews example, from http://en.wikinews.org/wiki/U.S._presidential_candidate_Gary_Johnson_leaves_GOP_to_vie_for_the_LP_nom

entity is represented by a URI starting with http://dbpedia.org/resource/ followed by the name of its associated Wikipedia title. See the next section for some example triples.

Pilot. While creating unambiguous descriptions is the NLG task known as referring expression generation, its NLU counterpart is anaphora resolution. We took a hand-annotated corpus for training anaphora resolution algorithms (Hasler et al., 2006) consisting of 74 documents containing 239 coreference chains. Each of the chains is an entity that can be used for our experiments, if the entity is in dbPedia and there are other suitable distractors in the same document. We hand annotated each of those 239 coreference chains by type (person, organization and location) and associated them to dbPedia URIs for the ones we found on Wikipedia. We found roughly half of the chains in dbPedia (106 out of 239, 44%). This percentage speaks of the coverage of dbPedia for OD REG. However, only 16 documents contain multiple entities of the same type and present in dbPedia, our pilot study criteria. These 16 documents result in the 16 tasks for our pilot. For a large scale evaluation we turned to Wikinews.

Wikinews. Wikinews is a news service operated as a wiki. As the news articles are interspersed with *interwiki* links, multiple entities can be disambiguated as Wikipedia pages (which in turn are dbPedia URIs). For example, in Figure 1, both the Libertarian Party and Republican Party can be considered potential distractors, as both are organizations.

The Wikimedia Foundation makes a database dump available for all Wikinews interwiki links (the links in braces in the above example). If a page contains more than one organization or person, we extracted the whole set of people (or organizations) as a referring expression task. To see whether a URI is a person or an organization we check for a birth date or creation date, respectively. In this manner, we obtained 4,230 tasks for people and 12,998 for organizations. This is dataset is freely available.[9]

4 Results

Pilot. The 16 tasks were split into 40 runs (a task spans n runs each, where n is the number of entities in the task, by rotating through the different alternative pairs of referent / set of distractors). From these tasks, Dale and Reiter produced no output 12 times and FB Brevity was unable to produce a result in 23 times. Gardent produced output for every run. We consider this an example of the increased expressive power of negative descriptions (it included a negation in 25% of the runs). For the other two algorithms, the lack of an unique triple differentiating one entity from the set of distractors seemed to be the main issue but there were multiple cases were FB ran out of memory for its queue of candidate nodes.

With respect to execution timings, Dale and Reiter ran into some corner cases and took time comparable to Gardent's algorithm. FB was 16 times slower (we found this counter-intuitive, as Gardent's algorithm is more demanding). Therefore, two of these algorithms were able to produce results using large scale ontological information. As FB ran into problems both in terms of execution time and failure rates, we omitted it from the large scale experiments.

We adjusted the parameters for the algorithms on this set to obtain the best possible quality output given the data and the problem. As such, we do not report quality assessments on the pilot data.

Wikinews. The tasks obtained from wikinews contained a large number of entities per task (an average of 12 people per task) and therefore span a large number of runs: 17,814 runs for people (from 4,230 tasks) and 44,080 for organizations (from 12,998 tasks).

On these large runs, execution time differences are in line with our *a priori* expectations: the greedy approach of Dale and Reiter is very fast[10] with Gardent's more comprehensive search taking about 40 times more time. Dale and Reiter failure rate was

[9] http://www.cs.famaf.unc.edu.ar/~pduboue/data/ also mirrored at http://duboue.ca/data.

[10] Dale and Reiter takes less than 3' for the 44,080 runs for organizations in a 2.3 GHz machine.

Referent	Dale and Reiter Output	Gardent Output
EB	{ (EB *occupation* Software_Freedom_Law_Center) }	{ (EB *occupation* Software_Freedom_Law_Center) }
LL	{ (LL *birthPlace* United_States), (LL, *occupation* Harvard_Law_School) }	{ (LL *birthPlace* Rapid_City,_South_Dakota) }
LT	{ (LT *occupation* Software_engineer) }	{ (LT *nationality* Finnish_American) }

Figure 2: Example output for the task: {'Eben_Moglen' (EB), 'Lawrence_Lessig' (LL), 'Linus_Torvalds' (LT) }.

comparable or better than in the pilot (for organizations that are more mixed, it was slightly lower but for people it was as low 2.8%). Gardent missed 2% of the people (and only 54 organizations), employing negatives 14% of the time for people and 12% of the time for organizations.

Evaluating referring expressions is hard. 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 Open Domain case, a major problem is the need for the judges to be acquainted with some of the less popular entities in the training set. At this point in our research, we decided to analyze the quality of a sample of the output ourselves. This process involved consulting information about each entity to determine the soundness of the result.

We looked at a random sample of 20 runs and annotated it by two authors, measuring a Cohen's κ of 60% for annotating DD results and 79% for determining whether the folksonomy had enough information to build a satisfactory DD. We then extended the evaluation to 60 runs and annotated them by one author. We found that Dale and Reiter produced a satisfactory DD in 41.6% of the cases and Gardent in 43.4% of the cases and that the folksonomy contained enough information 81.6% of the time. Figure 2 shows some example output.

From the evaluation we learned that the default ordering strategy employed by Dale and Reiter is not stable across different types of people (compare: politicians vs. musicians) or organizations. We also saw that Gardent's algorithm in many cases selected a single triple with very little practical value (an obscure fact about the entity) or a negative piece of information which is actually true for the referent but it is a missing piece of information.

The first two problems can be solved by either further subdividing the taxonomies of entities or (more interestingly) by incorporating some measure about the salience of each piece of information, a possibility which we will discuss next. The last issue can be addressed by having some form of meaningful default value.

The negations produced by Gardent's algorithm highlighted errors on the folksonomy. For example, when referring to China with distractors Peru and Taiwan, it will produce "the place where they do not speak Chinese," as China has the different Chinese dialects spelled out on the folksonomy (and some Peruvians do speak Chinese). Given these limitations, we find the current results very encouraging and we believe folksonomies can help focus on robust NLG for noisy (ontological) inputs.

5 Discussion

We have shown that by using a folksonomy it should be possible to deploy traditional NLG referring expression generation algorithms in Open Domain tasks. To fulfill this vision, three tasks remain:

Dealing with missing information. Some form of *smart default values* are needed, we are considering using a nearest-neighbor approach to find ontological siblings which can provide such defaults.

Estimating salience of each piece of ontological information. The importance for each triple has to be obtained in a way consistent with the Open Domain nature of the task. For this problem, we believe search engine salience can be of great help.

Transform the extracted triples into actual text. This problem has received attention in the past. We would like to explore traditional surface realizer with a custom-made grammar.

Acknowledgments

We would like to thank the anonymous reviewers as well as Annie Ying and Victoria Reggiardo.

References

Siddharthan Advaith, Nenkova Ani, and McKeown Kathleen. 2011. Information status distinctions and referring expressions: An empirical study of references to people in news summaries. *Computational Linguistics*, 37(4):811–842.

Anja Belz, Eric Kow, Jette Viethen, and Albert Gatt. 2010. Generating referring expressions in context: The grec task evaluation challenges. In Emiel Krahmer and Marit Theune, editors, *Empirical Methods in Natural Language Generation*, volume 5790 of *Lecture Notes in Computer Science*, pages 294–327. Springer.

C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. 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.

B. Bohnet. 2007. is-fbn, is-fbs, is-iac: The adaptation of two classic algorithms for the generation of referring expressions in order to produce expressions like humans do. *MT Summit XI, UCNLG+ MT*, pages 84–86.

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

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

A. Gatt, I. Van Der Sluis, and K. Van Deemter. 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.

L. Hasler, C. Orasan, and K. Naumann. 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.

Kathleen R. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, J. L. Klavans, A. Nenkova, C. Sable, B. Schiffman, and S. Sigelman. 2002. Tracking and summarizing news on a daily basis with columbia's newsblaster. In *Proc. of HLT*.

Kishore Papinini, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. Technical report, IBM.