Random Walk Factoid Annotation for Collective Discourse

Ben King **Rahul Jha**
Department of EECS
University of Michigan
Ann Arbor, MI
benking@umich.edu
rahuljha@umich.edu

Dragomir R. Radev
Department of EECS
School of Information
University of Michigan
Ann Arbor, MI
radev@umich.edu

Robert Mankoff *
The New Yorker Magazine
New York, NY
bob_mankoff
@newyorker.com

Abstract

In this paper, we study the problem of automatically annotating the factoids present in collective discourse. Factoids are information units that are shared between instances of collective discourse and may have many different ways of being realized in words. Our approach divides this problem into two steps, using a graph-based approach for each step: (1) factoid discovery, finding groups of words that correspond to the same factoid, and (2) factoid assignment, using these groups of words to mark collective discourse units that contain the respective factoids. We study this on two novel data sets: the New Yorker caption contest data set, and the crossword clues data set.

1 Introduction

Collective discourse tends to contain relatively few *factoids*, or information units about which the author speaks, but many *nuggets*, different ways to speak about or refer to a factoid (Qazvinian and Radev, 2011). Many natural language applications could be improved with good factoid annotation.

Our approach in this paper divides this problem into two subtasks: discovery of factoids, and assignment of factoids. We take a graph-based approach to the problem, clustering a word graph to discover factoids and using random walks to assign factoids to discourse units.

We also introduce two new datasets in this paper, covered in more detail in section 3. The New Yorker cartoon caption dataset, provided by Robert Mankoff, the cartoon editor at The New Yorker magazine, is composed of reader-submitted captions for a cartoon published in the magazine. The crossword clue dataset consists of word-clue pairs used in major American crossword puzzles, with most words having several hundred different clues published for it.

Figure 1: The cartoon used for the New Yorker caption contest #331.

The term "factoid" is used as in (Van Halteren and Teufel, 2003), but in a slightly more abstract sense in this paper, denoting a set of related words that should ideally refer to a real-world entity, but may not for some of the less coherent factoids. The factoids discovered using this method don't necessarily correspond to the factoids that might be chosen by annotators.

For example, given two user-submitted cartoon captions

- "When they said, 'Take us to your leader,' I don't think they meant your mother's house,"
- and "You'd better call your mother and tell her to set a few extra place settings,"

a human may say that they share the factoid called "mother." The automatic methods however, might say that these captions share *factoid3*, which is identified by the words "mother," "in-laws," "family," "house," etc.

The layout of this paper is as follows: we review related work in section 2, we introduce the datasets

*Cartoon Editor, The New Yorker magazine

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 249–254,
Sofia, Bulgaria, August 4-9 2013.

in detail in section 3, we describe our methods in section 4, and report results in section 5.

2 Related Work

The distribution of factoids present in text collections is important for several NLP tasks such as summarization. The Pyramid Evaluation method (Nenkova and Passonneau, 2004) for automatic summary evaluation depends on finding and annotating factoids in input sentences. Qazvinian and Radev (2011) also studied the properties of factoids present in collective human datasets and used it to create a summarization system. Hennig et al. (2010) describe an approach for automatically learning factoids for pyramid evaluation using a topic modeling approach.

Our random-walk annotation technique is similar to the one used in (Hassan and Radev, 2010) to identify the semantic polarity of words. Das and Petrov (2011) also introduced a graph-based method for part-of-speech tagging in which edge weights are based on feature vectors similarity, which is like the corpus-based lexical similarity graph that we construct.

3 Data Sets

We introduce two new data sets in this paper, the New Yorker caption contest data set, and the crossword clues data set. Though these two data sets are quite different, they share a few important characteristics. First, the discourse units tend to be short, approximately ten words for cartoon captions and approximately three words for crossword clues. Second, though the authors act independently, they tend to produce surprisingly similar text, making the same sorts of jokes, or referring to words in the same sorts of ways. Thirdly, the authors often try to be non-obvious: obvious jokes are often not funny, and obvious crossword clues make a puzzle less challenging.

3.1 New Yorker Caption Contest Data Set

The New Yorker magazine holds a weekly contest[1] in which they publish a cartoon without a caption and solicit caption suggestions from their readers. The three funniest captions are selected by the editor and published in the following weeks. Figure 1 shows an example of such a cartoon, while Table 1 shows examples of captions, including its winning captions. As part of this research project, we have acquired five cartoons along with all of the captions submitted in the corresponding contest.

I don't care what planet they are from, they can pass on the left like everyone else.
I don't care what planet they're from, they should have the common courtesy to dim their lights.
I don't care where he's from, you pass on the left.
If he wants to pass, he can use the right lane like everyone else.
When they said, 'Take us to your leader,' I don't think they meant your mother's house.
They may be disappointed when they learn that "our leader" is your mother.
You'd better call your mother and tell her to set a few extra place settings.
If they ask for our leader, is it Obama or your mother?
Which finger do I use for aliens?
I guess the middle finger means the same thing to them.
I sense somehow that flipping the bird was lost on them.
What's the Klingon gesture for "Go around us, jerk?"

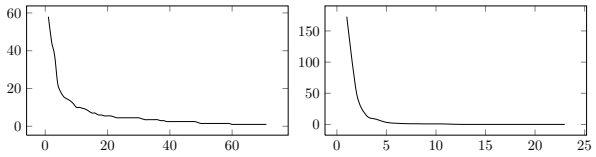
Table 1: Captions for contest #331. Finalists are listed in italics.

While the task of automatically identifying the funny captions would be quite useful, it is well beyond the current state of the art in NLP. A much more manageable task, and one that is quite important for the contest's editor is to annotate captions according to their factoids. This allows the organizers of the contest to find the most frequently mentioned factoids and select representative captions for each factoid.

On average, each cartoon has 5,400 submitted captions, but for each of five cartoons, we sampled 500 captions for annotation. The annotators were instructed to mark factoids by identifying and grouping events, objects, and themes present in the captions, creating a unique name for each factoid, and marking the captions that contain each factoid. One caption could be given many different labels. For example, in cartoon #331, such factoids may be "bad directions", "police", "take me to your leader", "racism", or "headlights". After annotating, each set of captions contained about 60 factoids on average. On average a caption was annotated with 0.90 factoids, with approximately 80% of the discourse units having at least one factoid, 20% having at least two, and only 2% having more than two. Inter-annotator agreement was moderate, with an F1-score (described more in section 5) of 0.6 between annotators.

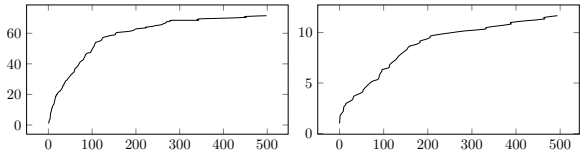
As van Halteren and Teufel (2003) also found

[1] http://www.newyorker.com/humor/caption



(a) (b)

Figure 2: Average factoid frequency distributions for cartoon captions (a) and crossword clues (b).



(a) (b)

Figure 3: Growth of the number of unique factoids as the size of the corpus grows for cartoon captions (a) and crossword clues (b).

when examining factoid distributions in human-produced summaries, we found that the distribution of factoids in the caption set for each cartoon seems to follow a power law. Figure 2 shows the average frequencies of factoids, when ordered from most- to least-frequent. We also found a Heap's law-type effect in the number of unique factoids compared to the size of the corpus, as in Figure 3.

3.2 Crossword Clues Data Set

Clues in crossword puzzles are typically obscure, requiring the reader to recognize double meanings or puns, which leads to a great deal of diversity. These clues can also refer to one or more of many different senses of the word. Table 2 shows examples of many different clues for the word "tea". This table clearly illustrates the difference between factoids (the senses being referred to) and nuggets (the realization of the factoids).

The website `crosswordtracker.com` collects a large number of clues that appear in different published crossword puzzles and aggregates them according to their answer. From this site, we collected 200 sets of clues for common crossword answers.

We manually annotated 20 sets of crossword clues according to their factoids in the same fashion as described in section 3.1. On average each set of clues contains 283 clues and 15 different factoids. Inter-annotator agreement on this dataset was quite high with an F1-score of 0.96.

Clue	Sense
Major Indian export	drink
Leaves for a break?	drink
Darjeeling, e.g.	drink
Afternoon social	event
4:00 gathering	event
Sympathy partner	film
Mythical Irish queen	person
____ Party movement	political movement
Word with rose or garden	plant and place

Table 2: Examples of crossword clues and their different senses for the word "tea".

4 Methods

4.1 Random Walk Method

We take a graph-based approach to the discovery of factoids, clustering a word similarity graph and taking the resulting clusters to be the factoids. Two different graphs, a word co-occurrence graph and a lexical similarity graph learned from the corpus, are compared. We also compare the graph-based methods against baselines of clustering and topic modeling.

4.1.1 Word Co-occurrence Graph

To create the word co-occurrence graph, we create a link between every pair of words with an edge weight proportional to the number of times they both occur in the same discourse unit.

4.1.2 Corpus-based Lexical Similarity Graph

To build the lexical similarity graph, a lexical similarity function is learned from the corpus, that is, from one set of captions or clues. We do this by computing feature vectors for each lemma and using the cosine similarity between these feature vectors as a lexical similarity function. We construct a word graph with edge weights proportional to the learned similarity of the respective word pairs.

We use three types of features in these feature vectors: context word features, context part-of-speech features, and spelling features. Context features are the presence of each word in a window of five words (two words on each side plus the word in question). Context part-of-speech features are the part-of-speech labels given by the Stanford POS tagger (Toutanova et al., 2003) within the same window. Spelling features are the counts of all character trigrams present in the word.

Table 3 shows examples of similar word pairs from the set of crossword clues for "tea". From

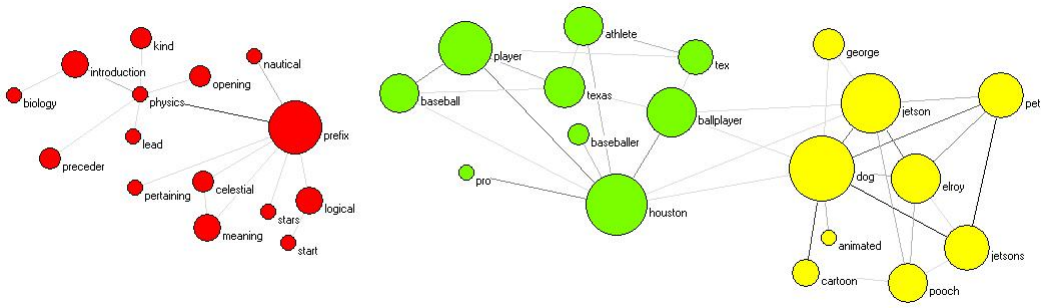


Figure 4: Example of natural clusters in a subsection of the word co-occurrence graph for the crossword clue "astro".

Word pair	Sim.
(white-gloves, white-glove)	0.74
(may, can)	0.57
(midafternoon, mid-afternoon)	0.55
(company, co.)	0.46
(supermarket, market)	0.53
(pick-me-up, perk-me-up)	0.44
(green, black)	0.44
(lady, earl)	0.39
(kenyan, indian)	0.38

Table 3: Examples of similar pairs of words as calculated on the set of crossword clues for "tea".

this table, we can see that this method is able to successfully identify several similar word pairs that would be missed by most lexical databases: minor lexical variations, such as "pick-me-up" vs. "perk-me-up"; abbreviations, such as "company" and "co."; and words that are similar only in this context, such as "lady" and "earl" (referring to Lady Grey and Earl Grey tea).

4.1.3 Graph Clustering

To cluster the word similarity graph, we use the Louvain graph clustering method (Blondel et al., 2008), a hierarchical method that optimizes graph modularity. This method produces several hierarchical cluster levels. We use the highest level, corresponding to the fewest number of clusters.

Figure 4 shows an example of clusters found in the word graph for the crossword clue "astro". There are three obvious clusters, one for the Houston Astros baseball team, one for the dog in the Jetsons cartoon, and one for the lexical prefix "astro-". In this example, two of the clusters are connected by a clue that mentions multiple senses, "Houston ballplayer or Jetson dog".

4.1.4 Random Walk Factoid Assignment

After discovering factoids, the remaining task is to annotate captions according to the factoids they contain. We approach this problem by taking random walks on the word graph constructed in the previous sections, starting the random walks from words in the caption and measuring the hitting times to different clusters.

For each discourse unit, we repeatedly sample words from it and take Markov random walks starting from the nodes corresponding to the selected and lasting 10 steps (which is enough to ensure that every node in the graph can be reached). After 1000 random walks, we measure the average hitting time to each cluster, where a cluster is considered to be reached by the random walk the first time a node in that cluster is reached. Heuristically, 1000 random walks was more than enough to ensure that the factoid distribution had stabilized in development data.

The labels that are applied to a caption are the labels of the clusters that have a sufficiently low hitting time. We perform five-fold cross validation on each caption or set of clues and tune the threshold on the hitting time such that the average number of labels per unit produced matches the average number of labels per unit in the gold annotation of the held-out portion.

For example, a certain caption may have the following hitting times to the different factoid clusters:

factoid1	0.11
factoid2	0.75
factoid3	1.14
factoid4	2.41

If the held-out portion has 1.2 factoids per caption, it may be determined that the optimal thresh-

old on the hitting times is 0.8, that is, a threshold of 0.8 produces 1.2 factoids per caption in the test-set on average. In this case *factoid1* and *factoid2* would be marked for this caption, since the hitting times fall below the threshold.

4.2 Clustering

A simple baseline that can act as a surrogate for factoid annotation is clustering of discourse units, which is equivalent to assigning exactly one factoid (the name of its cluster) to each discourse unit. As our clustering method, we use C-Lexrank (Qazvinian and Radev, 2008), a method that has been well-tested on collective discourse.

4.3 Topic Model

Topic modeling is a natural way to approach the problem of factoid annotation, if we consider the topics to be factoids. We use the Mallet (McCallum, 2002) implementation of Latent Dirichlet Allocation (LDA) (Blei et al., 2003). As with the random walk method, we perform five-fold cross validation, tuning the threshold for the average number of labels per discourse unit to match the average number of labels in the held-out portion. Because LDA needs to know the number of topics *a priori*, we set the number of topics to be equal to the true number of factoids. We also use the average number of unique factoids in the held-out portion as the number of LDA topics.

5 Evaluation and Results

We evaluate this task in a way similar to pairwise clustering evaluation methods, where every pair of discourse units that should share at least one factoid and does is a true positive instance, every pair that should share a factoid and does not is a false negative, etc. From this we are able to calculate precision, recall, and F1-score. This is a reasonable evaluation method, since the average number of factoids per discourse unit is close to one. Because the factoids discovered by this method don't necessarily match the factoids chosen by the annotators, it doesn't make sense to try to measure whether two discourse units share the "correct" factoid.

Tables 4 and 5 show the results of the various methods on the cartoon captions and crossword clues datasets, respectively. On the crossword clues datasets, the random-walk-based methods are clearly superior to the other methods tested, whereas simple clustering is more effective on the cartoon captions dataset.

Method	Prec.	Rec.	F1
LDA	0.318	0.070	0.115
C-Lexrank	0.131	0.347	0.183
Word co-occurrence graph	0.115	0.348	0.166
Word similarity graph	0.093	0.669	0.162

Table 4: Performance of various methods annotating factoids for cartoon captions.

Method	Prec.	Rec.	F1
LDA	0.315	0.067	0.106
C-Lexrank	0.702	0.251	0.336
Word co-occurrence graph	0.649	0.257	0.347
Word similarity graph	0.575	0.397	0.447

Table 5: Performance of various methods annotating factoids for crossword clues.

In some sense, the two datasets in this paper both represent difficult domains, ones in which authors are intentionally obscure. The good results acheived on the crossword clues dataset indicate that this obscurity can be overcome when discourse units are short. Future work in this vein includes applying these methods to domains, such as newswire, that are more typical for summarization, and if necessary, investigating how these methods can best be applied to domains with longer sentences.

References

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 600–609.

Ahmed Hassan and Dragomir Radev. 2010. Identifying text polarity using random walks. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 395–403. Association for Computational Linguistics.

Leonhard Hennig, Ernesto William De Luca, and Sahin Albayrak. 2010. Learning summary content units with topic modeling. In *Proceedings of the 23rd*

International Conference on Computational Linguistics: Posters, COLING '10, pages 391–399, Stroudsburg, PA, USA. Association for Computational Linguistics.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method.

Vahed Qazvinian and Dragomir R Radev. 2008. Scientific paper summarization using citation summary networks. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 689–696. Association for Computational Linguistics.

Vahed Qazvinian and Dragomir R Radev. 2011. Learning from collective human behavior to introduce diversity in lexical choice. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language techologies*, pages 1098–1108.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics.

Hans Van Halteren and Simone Teufel. 2003. Examining the consensus between human summaries: initial experiments with factoid analysis. In *Proceedings of the HLT-NAACL 03 on Text summarization workshop-Volume 5*, pages 57–64. Association for Computational Linguistics.