Topics in Algorithmic Data Analysis
Assignment 4: Connecting the dots

Muhammad Yaseen - 2577833
s8muyase@stud.uni-saarland.de

Introduction

In this report we take a look at three different problems: First one deals with forming a coherent narrative within the news domain [1]. Second work focuses on developing a computational model of humor with the aim of assigning a humor score to cartoon captions [2]. Lastly, we look at the problem of analogy mining [3] in product design context.

Our goal is to explore the connections between these disparate works while also evaluating them individually. In service of brevity, we introduce helpful abbreviations for these three works, and use: CTD for "Connecting the Dots" [1], IJ for "Inside Jokes" [2], and AM for "Analogy Mining" [3].

1 Connecting the Dots

In brief, idea of this work can be described as: Given two news articles, find a coherent chain of news articles which links the two endpoints while respecting chronological and coherence constraints. The motivation behind this is to present a reasonably complete narrative and context so that users can effectively understand not only the story but also how it developed or progressed over time without having to look at potentially irrelevant articles when they go about this task themselves. Interestingly, this can be considered an automation of a similar practice in Open Source Intelligence (OSINT) community [1] which does this task not just with news stories but also includes other websites and social media information.

As a measure of **coherence** they formalize **influence** of words on adjacent articles in terms of word activation patterns, which is without precedence. Many existing established techniques could have been used to measure inter-document coherence e.g Document Vector, Sentence Embeddings, Topic Models, Latent Semantic Analysis etc. Many other important notions like semantic relatedness and co-occurence could also have been exploited. They motivated this formalization by some emperical observations on their corpus which do not convincingly justify this ad-hoc measure, specially when other addressing these methods are already available.

[1] Thread on recent Sino-Indian border conflict: https://twitter.com/Natsecjeff/status/1272792219515326464

As for their ILP constraints which aim to reduce redundancy, we note that the problem of retrieval system results diversification is well-studied in IR in fact, there exist several methods which address both **diversity** and **novelty** of the results in context of polysemy (same word, different meanings) and synonymy (same meaning, different words) as well as **maximum marginal relevance** formalism to address novelty. Similarly, the two modifications they propose to the generated chains which incorporate user feedback i.e. **refinement** and **term importance** can be mapped to **relevance feedback approach** in IR [4].

Above points notwithstanding, we examine the methodology and user study itself closely. They distinguish between local and global coherence i.e. an article in chain must be related to all other articles in the chain and pairwise relevance is not enough as it might lead to a lot of erratic transitions. However, their chosen examples do not take into account the difference between long-running events like 2008 Financial Crash, relatively short-lived events like recent Sino-Indian border conflict, and medium term events like Brexit. They will have considerably different challenge because of amount of information, frequency of information, and diversity in keywords.

In the user study, they measure a user's topic **familiarity** on scale of 1 to 5 but then measure the three properties of chains i.e. **relevance**,**coherence**, **redundancy** in binary. Alternatively, to reduce arbitrariness in familiarity rating they could have done a pre and post chain reading questionnaire with questions derived from the topic. Both question formulation and answer verification can be easily automated with entity recognition and template methods. Then, a natural measure would be difference of correct answers in pre and post questionnaire. Similarly, we believe that relevance and redundancy could have been gauged better if they had asked workers to also point out which articles are redundant (resp. relevant) in a chain. This would provide more fine grained data than a binary judgement.

In conclusion, we hold that the problem could have been cast in the framework of NLP and IR directly without the scaffolding of and Integer Linear Programming algorithm. It would also be interesting to see if it can be extended to more than just news coverage, as mentioned in OSINT example.

2 Inside Jokes

Central idea behind this work is to develop a computational model of humor. The authors focus their task on ranking cartoon captions submitted to New Yorker magazine i.e. giving them a humor score. Their goal is to get insights as to what makes a caption funnier than the others and thus potentially automate the process of selecting the best cartoon caption. Thus, the final task can be considered as modeling the judgement process of New Yorker judges and readership as they rate funnier captions.

First they turn the captions into a vector representation based on language model, readability, syntactic features. Then in a somewhat artificial manner they concatenate features of both jokes i.e. the one judged as funnier and the less-funny one and their differences with label to turn it into classification problem. A more natural **pairwise comparison** paradigm exists in IR, which could have been used.

In the part where they gathered pairwise ratings on jokes from AMT workers, it might have been beneficial to also get annotations from users selecting words / phrases which make the joke they selected more funny compared to other. Similar to the annotations used in analogy mining (discussed later) to highlight purpose and mechanism in a description. It gives fine grained information and intersection of such phrases could have provided more concrete data and probably would have eliminated the arbitrariness of using perplexity of LMs, readability scores as a proxy for those. With these annotations they could have modeled the problem as predicting the funny parts in a give joke and then score them.

In the final tournament under limitations section they provided only two examples of captions which the judges ranked high but their method didn't. Given that the percentage coverage of their method is not very high it would have been more insightful to examine closely other failed examples as well, while possibly also categorizing the reasons for which the method fails. It would have also been more relevant to their original goal of modeling humor. Failure cases are important to see what the model lacks.

Lastly, we note that cartoon images follows a very specific pattern i.e. something unusual in a rather mundane setting e.g. a car dealership scene with a car having bear-like feet instead of regular tires. This further limits the kind of humorous content readers might be primed to produce from this visual stimulus. This, combine with the fact that they didn't include a detailed study of either success or failure cases, but rather focused on classification accuracy leads us to conclude that this work doesn't seem to live up to the audacious goal of developing a computational humor model.

3 Analogy Mining

Principal motivation for this work is to find creative analogies in product design domain. Authors use crowd-sourced annotations on unstructured product description data to learn **purpose** and **mechanism** vectors for each product. They situate their work in distinction to structured analogy mining e.g. where one could run an SQL-like query to find items satisfying a predefined set of conditions. This however requires existence of a structured dataset for analogy mining which is hard to find.

To circumvent this, they acquire annotations via crowdsourcing on product description texts with words and phrases labelled as describing either product purpose or mechanism. These purpose (resp. mechanism) annotations are then combined into a unified product purpose (resp. product mechanism) representation via TF-IDF weighted GloVe embeddings. They learn to map product description w_i representation to its purpose and mechanism embeddings (p_i, m_i) via $f(w_i)$. This representation can be used to do queries like near-purpose-far-mechanism, or clustering by purpose then finding different mechanism etc.

To collect analogies i.e. product pairs that are in some sense analogous they use crowdsourcing. They set a target doc and then based on KW search find other documents related to it and then ask the workers to annotate them as match (positive examples) if they consider it an analogy. Pairs not matched are implicitly treat as negative examples i.e. non-analogies. Idea is that KW search retrieves only relevant docs which may or

may not be analogous and workers are used to filter them down to analogous pairs. It is interesting to note that workers didn't seem to have a problem differentiating between purpose and mechanism whereas they had trouble grasping the problem of analogy. We believe that they could have resolved this at the cost of time. For example, by requiring the workers to also annotate upto k pairs (to keep it manageable) pairs of sentences or phrases in target doc and analogies. This would help them explain why there was disagreement in the first place and also similar to what was proposed in Inside Jokes score documents based on these more fine grained annotations.

Finally, we note that diversifying the inspiration set based on mechanism i.e. products with various different mechanism but same purpose are preferred can be treated as simple IR result set diversification problem given their representations.

Thus in our view the key novelty of this work is introducing a domain-specific document representation i.e. purpose and mechanism vectors and dataset collection. The rest can be considered a specific application of the more general NLP and IR machinery.

4 Overall thematic similarities

In all three works, ideas from Natural Language Processing (NLP) and Information Retrieval (IR) domain are used either as guiding principles, as baselines, defaults, or both. Compared to IJ, we find the goal in CTD and AM to be more clearly defined and final user studies are more directly relevant to problem definition. However, in both IJ and AM a large portion of acquired data is discarded because of non-agreement in the rankings/matches. Author chalk it up to either the inherent subjectivity or difficulty of the problem. But as we discussed in the respective sections, it could have been avoided by a slight reformulation of problem. We summarize the thematic similarities in the following table:

	Inside Jokes	Connecting the Dots	Analogy Mining
Data representation intuition	language models and gumor research	purpose and mechanism bifurcation in engineering machine design	inter-document coherence
Dichotomies	same joke, different delivery	near purpose, far mechanism	same overall narrative, different detail
Relevance Judgements	funny / less funny	analogy / non-analogy	coherence and relatedness
Crowdsourcing judgements	pairwise joke rank		purpose mechanism annotations and pairwise analogies judgements
User study Criteria	expert editor judgements	graduate student judgements based on given subjective criteria	pairwise chain comparison in user study

5 A comment on evaluation metrics and reproducibility

The problems addressed in the considered works i.e. coherence, humor, analogies are inherently subjective and don't lend themselves easily to any formal definition. Thus authors had to define indirect measures e.g. word activation, unusual language, purpose-mechanism vectors and relate them to problem at hand for any kind of analysis or algorithm. As a consequence, these problems also lack any direct well-defined numerical metric like classification accuracy or mean squared error etc. to judge the final result. For this reason they relied on human judgements via user experience studies for final validation.

This however, raises another issue of reproducibility. Even if we consider the dataset / code to be available, replicating the results will depend on relying on human judgements. Secondly, the measures on which these judgements were obtained are also themselves rather coarse and susceptible to variations as explained. These two points combined cast doubt on how other works can be compared to these works in future.

References

[1] D. Shahaf and C Guestrin. "Connecting the dots between news articles". In: *In Proceedings of the 16th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)* (2010), pp. 623–632.

[2] D. Shahaf, E. Horvitz, and R Mankoff. "Inside Jokes: Identifying Humorous Cartoon Captions". In: *In Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)* (2015), pp. 1065–1074.

[3] T. Hope et al. "Accelerating Innovation Through Analogy Mining". In: *In Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)* (2017), pp. 235–243.

[4] Christopher Manning, Hinrich Schütze, and Prabhakar Raghavan. *Introduction to Information Retrieval*. Cambridge University Press, 2008. ISBN: 9780521865715.