An Entity-Focused Approach to Generating Company Descriptions

Gavin Saldanha* Columbia University
Or Biran** Columbia University
Kathleen McKeown** Columbia University
Alfio Gliozzo† IBM Watson

* gvs2106@columbia.edu ** {orb, kathy}@cs.columbia.edu
†gliozzo@us.ibm.com

Abstract

Finding quality descriptions on the web, such as those found in Wikipedia articles, of newer companies can be difficult: search engines show many pages with varying relevance, while multi-document summarization algorithms find it difficult to distinguish between core facts and other information such as news stories. In this paper, we propose an entity-focused, hybrid generation approach to automatically produce descriptions of previously unseen companies, and show that it outperforms a strong summarization baseline.

1 Introduction

As new companies form and grow, it is important for potential investors, procurement departments, and business partners to have access to a 360-degree view describing them. The number of companies worldwide is very large and, for the vast majority, not much information is available in sources like Wikipedia. Often, only firmographics data (e.g. industry classification, location, size, and so on) is available. This creates a need for cognitive systems able to aggregate and filter the information available on the web and in news, databases, and other sources. Providing good quality natural language descriptions of companies allows for easier access to the data, for example in the context of virtual agents or with text-to-speech applications.

In this paper, we propose an entity-focused system using a combination of targeted (knowledge base driven) and data-driven generation to create company descriptions in the style of Wikipedia descriptions. The system generates sentences from RDF triples, such as those found in DBPedia and Freebase, about a given company and combines these with sentences on the web that match learned expressions of relationships. We evaluate our hybrid approach and compare it with a targeted-only approach and a data-driven-only approach, as well as a strong multi-document summarization baseline. Our results show that the hybrid approach performs significantly better than either approach alone as well as the baseline.

The targeted (TD) approach to company description uses Wikipedia descriptions as a model for generation. It learns how to realize RDF relations that have the company as their subject: each relation contains a company/entity pair and it is these pairs that drive both content and expression of the company description. For each company/entity pair, the system finds all the ways in which similar company/entity pairs are expressed in other Wikipedia company descriptions, clustering together sentences that express the same company/entity relation pairs. It generates templates for the sentences in each cluster, replacing the mentions of companies and entities with typed slots and generates a new description by inserting expressions for the given company and entity in the slots. All possible sentences are generated from the templates in the cluster, the resulting sentences are ranked and the best sentence for each relation selected to produce the final description. Thus, the TD approach is a top-down approach, driven to generate sentences expressing the relations found in the company's RDF data using realizations that are typically used on Wikipedia.

In contrast, the data-driven (DD) approach uses a semi-supervised method to select sentences from descriptions about the given company on the web. Like the TD approach, it also begins with a seed set of relations present in a few companies' DBPedia entries, represented as company/entity pairs, but instead of looking at the corresponding Wikipedia articles, it learns patterns that are typ-

ically used to express the relations on the web. In the process, it uses bootstrapping (Agichtein and Gravano, 2000) to learn new ways of expressing the relations corresponding to each company/entity pair, alternating with learning new pairs that match the learned expression patterns. Since the bootstrapping process is driven only by company/entity pairs and lexical patterns, it has the potential to learn a wider variety of expressions for each pair and to learn new relations that may exist for each pair. Thus, this approach lets data for company descriptions on the web determine the possible relations and patterns for expressing those relations in a bottom-up fashion. It then uses the learned patterns to select matching sentences from the web about a target company.

2 Related Work

The TD approach falls into the generation pipeline paradigm (Reiter and Dale, 1997), with content selection determined by the relation in the company's DBpedia entry while microplanning and realization are carried out through template generation. While some generation systems, particularly in early years, used sophisticated grammars for realization (Matthiessen and Bateman, 1991; Elhadad, 1991; White, 2014), in recent years, template-based generation has shown a resurgence. In some cases, authors focus on document planning and sentences in the domain are stylized enough that templates suffice (Elhadad and Mckeown, 2001; Bouayad-Agha et al., 2011; Gkatzia et al., 2014; Biran and McKeown, 2015). In other cases, learned models that align database records with text snippets and then abstract out specific fields to form templates have proven successful for the generation of various domains (Angeli et al., 2010; Kondadadi et al., 2013). Others, like us, target atomic events (e.g., date of birth, occupation) for inclusion in biographies (Filatova and Prager, 2005) but the templates used in other work are manually encoded.

Sentence selection has also been used for question answering and query-focused summarization. Some approaches focus on selection of relevant sentences using probabilistic approaches (Daumé III and Marcu, 2005; Conroy et al., 2006), semi-supervised learning (Wang et al., 2011) and graph-based methods (Erkan and Radev, 2004; Otterbacher et al., 2005). Yet others use a mixture of targeted and data-driven methods for a pure sentence selection system (Blair-Goldensohn et al., 2003; Weischedel et al., 2004; Schiffman et al., 2001). In our approach, we target both relevance and variety of expression, driving content by selecting sentences that match company/entity pairs and inducing multiple patterns of expression. Sentence selection has also been used in prior work on generating Wikipedia overall articles (Sauper and Barzilay, 2009). Their focus is more on learning domain-specific templates that control the topic structure of an overview, a much longer text than we generate.

3 Targeted Generation

The TD system uses a development set of 100 S&P500 companies along with their Wikipedia articles and DBPedia entries to form templates. For each RDF relation with the company as the subject, it identifies all sentences in the corresponding article containing the entities in the relation. The specific entities are then replaced with their relation to create a template. For example, "Microsoft was founded by Bill Gates and Paul Allen" is converted to "⟨company⟩ was founded by ⟨founder⟩," with conjoined entities collapsed into one slot. Many possible templates are created, some of which contain multiple relations (e.g., "⟨company⟩, located in ⟨location⟩, was founded by ⟨founder⟩"). In this way the system learns how Wikipedia articles express relations between the company and its key entities (founders, headquarters, products, etc).

At generation time, we fill the template slots with the corresponding information from the RDF entries of the target company. Conjunctions are inserted when slots are filled by multiple entities. Continuing with our example, we might now produce the sentence "Palantir was founded by Peter Thiel, Alex Karp, Joe Lonsdale, Stephen Cohen, and Nathan Gettings" for target company Palantir. Preliminary results showed that this method was not adequate - the data for the target company often lacked some of the entities needed to fill the templates. Without those entities the sentence could not be generated. As Wikipedia sentences tend to have multiple relations each (high information density), many sentences containing important, relevant facts were discarded due to phrases that mentioned lesser facts we did not have the data to replace. We therefore added a post-processing step to remove, if possible, any phrases

from the sentence that could not be filled; otherwise, the sentence is discarded.

This process yields many potential sentences for each relation, of which we only want to choose the best. We cluster the newly generated sentences by relation and score each cluster. Sentences are scored according to how much information about the target company they contain (number of replaced relations). Shorter sentences are also weighted more as they are less likely to contain extraneous information, and sentences with more post-processing are scored lower. The highest scored sentence for each relation type is added to the description as those sentences are the most informative, relevant, and most likely to be grammatically correct.

4 Data-Driven Generation

The DD method produces descriptions using sentences taken from the web. Like the TD approach, it aims to produce sentences realizing relations between the input company and other entities. It uses a bootstrapping approach (Agichtein and Gravano, 2000) to learn patterns for expressing the relations. It starts with a seed set of company/entity pairs, representing a small subset of the desired relations, but unlike previous approaches, can generate additional relations as it goes.

Patterns are generated by reading text from the web and extracting those sentences which contain pairs in the seed set. The pair's entities are replaced with placeholder tags denoting the type of the entity, while the words around them form the pattern (the words between the tags are selected as well as words to the left and right of the tags). Each pattern thus has the form "⟨L⟩⟨T1⟩⟨M⟩⟨T2⟩⟨R⟩," where L, M, and R are respectively the words to the left of, between, and to the right of the entities. T1 is the type of the first entity, and T2 the type of the second. Like the TD algorithm, this is essentially a template based approach, but the templates in this case are not aligned to a relation between the entity and the company; only the type of entity (person, location, organization, etc) is captured by the tag.

New entity pairs are generated by matching the learned patterns against web text. A sentence is considered to match a pattern if it has the same entity types in the same order and its L, M, and R words fuzzy match the corresponding words in the pattern.[1] The entities are therefore assumed to be related since they are expressed in the same way as the seed pair. Unlike the TD approach, the actual relationship between the entities is unknown (since the only data we use is the web text, not the structured RDF data); all we need to know here is that a relationship exists.

We alternate learning the patterns and generating entity pairs over our development set of 100 companies. We then take all the learned patterns and find matching sentences in the Bing search results for each company in the set of target companies.[2] Sentences that match any of the patterns are selected and ranked by number of matches (more matches means greater probability of strong relation) before being added to the description.

4.1 Pruning and Ordering

After selecting the sentences for the description, we perform a post-processing step that removes noise and redundancy. To address redundancy, we remove those sentences which were conveyed previously in the description using exactly the same wording. Thus, sentences which are equal to or subsets of other sentences are removed. We also remove sentences that come from news stories; analysis of our results on the development set indicated that news stories rarely contain information that is relevant to a typical Wikipedia description. To do this we use regular expressions to capture common newswire patterns (e.g., [CITY, STATE: sentence]). Finally, we remove incomplete sentences ending in "...", which sometimes appear on websites which themselves contain summaries.

We order the selected sentences using a scoring method that rewards sentences based on how they refer to the company. Sentences that begin with the full company name get a starting score of 25, sentences that begin with a partial company name start with a score of 15, and sentences that do not contain the company name at all start at -15 (if they contain the company name in the middle of the sentece, they start at 0). Then, 10 points are added to the score for each *keyword* in the sentence (keywords were selected from the most populous DBPedia predicates where the subject is a company). This scoring algorithm was tuned on the development set. The final output is ordered in

[1] We use a threshold on the cosine similarity of the texts to determine whether they match.

[2] We excluded Wikipedia results to better simulate the case of companies which do not have a Wikipedia page

descending order of scores.

5 Hybrid system

In addition to the two approaches separately, we also generated hybrid output from a combination of the two. In this approach, we start with the DD output; if (after pruning) it has fewer than three sentences, we add the TD output and re-order.

The hybrid approach essentially supplements the large, more noisy web content of the DD output with the small, high-quality but less diverse TD output. For companies that are not consumer-facing or are relatively young, and thus have a relatively low web presence - our target population - this can significantly impact the description.

6 Experiments

To evaluate our approach, we compare the three versions of our output - generated by the TD, DD, and hybrid approach - against multi-document summaries generated (from the same search results used by our DD approach) by TextRank (Mihalcea and Tarau, 2004). For each one of the approaches as well as the baseline, we generated descriptions for all companies that were part of the S&P500 as of January 2016. We used our development set of 100 companies for tuning, and the evaluation results are based on the remaining 400.

We conducted two types of experiments. The first is an automated evaluation, where we use the METEOR score (Lavie and Agarwal, 2007) between the description generated by one of our approaches or by the baseline and the first section of the Wikipedia article for the company. In Wikipedia articles, the first section typically serves as an introduction or overview of the most important information about the company. METEOR scores capture the content overlap between the generated description and the Wikipedia text. To avoid bias from different text sizes, we set the same size limit for all descriptions when comparing them. We experimented with three settings: 150 words, 500 words, and no size limit.

In addition, we conducted a crowd-sourced evaluation on the CrowdFlower platform. In this evaluation, we presented human annotators with two descriptions for the same company, one described by our approach and one by the baseline, in random order. The annotators were then asked to choose which of the two descriptions is a better overview of the company in question (they were provided a link to the company's Wikipedia page for reference) and give a score on a 1-5 scale to each description. For quality assurance, each pair of descriptions was processed by three annotators, and we only included in the results instances where all three agreed. Those constituted 44% of the instances. In this evaluation we only used the *hybrid* version, and we limited the length of both the baseline and our output to 150 words to reduce bias from a difference in lengths and keep the descriptions reasonably short for the annotators.

	150 words	500 words	no limit
TextRank	13.7	12.8	6.3
DD	15.0	14.5	14.0
TD	11.3	11.3	11.3
Hybrid	15.5	14.6	14.2

Table 1: First experiment results: average METEOR scores for various size limits

	% best	Avg. score
TextRank	25.79	2.82
Hybrid	74.21	3.81

Table 2: Second experiment results: % of companies for which the approach was chosen as best by the human annotators, and average scores given

7 Results

The results of the automated evaluation are shown in Table 1. Our DD system achieves higher METEOR scores than the TextRank baseline under all size variations, while TD by itself is worse in most cases. In all cases the combined approach achieves a better result than the DD system by itself.

The results of the human evaluation are shown in Table 2. Here the advantage of our approach becomes much more visible: we clearly beat the baseline both in terms of how often the annotators chose our output to be better (almost 75% of the times) and in terms of the average score given to our descriptions (3.81 on a $1-5$ point scale).

All results are statistically significant, but the difference in magnitude between the results of the two experiments are striking: we believe that while the TextRank summarizer extracts sentences which are topically relevant and thus achieve results close to ours in terms of METEOR, the more structured entity-focused approach we present here is able to extract content that seems much more reasonable to humans as a general description. One example is shown in Figure 1.

Right from the start, we see that our system out-

performs TextRank. Our first sentence introduces the company and provides a critical piece of history about it, while TextRank does not even immediately name it. The hybrid generation output has a more structured output, going from the origins of the company via merger, to its board, and finally its products. TextRank's output, in comparison, focuses on the employee experience and only mentions products at the very end. Our system is much more suitable for a short description of the company for someone unfamilar with it.

TextRank: The company also emphasizes stretch assignments and on-the-job learning for development, while its formal training programs include a Masters in the Business of Activision(or "MBActivision") program that gives employees a deep look at company operations and how its games are made, from idea to development to store shelves. How easy is it to talk with managers and get the information I need? Will managers listen to my input? At Activision Blizzard, 78 percent of employees say they often or almost always experience a free and transparent exchange of ideas and information within the organization. Gaming is a part of day-to-day life at Activision Blizzard, and the company often organizes internal tournaments for Call of Duty, Hearthstone: Heroes of Warcraft, Destiny, Skylanders and other titles. What inspires employees' company spirit here Do people stand by their teams' work What impact do people have outside the organization.

Hybrid: Activision Blizzard was formed in 2007 from a merger between Activision and Vivendi Games (as well as Blizzard Entertainment, which had already been a division of Vivendi Games.) Upon merger, Activision Blizzard's board of directors initially formed of eleven members: six directors designated by Vivendi, two Activision management directors and three independent directors who currently serve on Activision's board of directors. It's comprised of Blizzard Entertainment, best known for blockbuster hits including World of Warcraft, Hearthstone: Heroes of Warcraft, and the Warcraft, StarCraft, and Diablo franchises, and Activision Publishing, whose development studios (including Infinity Ward, Toys for Bob, Sledgehammer Games, and Treyarch, to name just a few) create blockbusters like Call of Duty, Skylanders, Guitar Hero, and Destiny.

Figure 1: Descriptions for Activision Blizzard

8 Conclusion

We described two approaches to generating company descriptions as well as a hybrid approach. We showed that our output is overwhelmingly preferred by human readers, and is more similar to Wikipedia introductions, than the output of a state-of-the-art summarization algorithm.

These complementary methods each have their advantages and disadvantages: the TD approach ensures that typical expressions in Wikipedia company descriptions - known to be about the fundamental relations of a company - will occur in the generated output. However, since it modifies them, it risks generating ungrammatical sentences or sentences which contain information about another company. The latter can occur because the sentence is uniquely tied to the original. For instance, the following Wikipedia sentence fragment – "Microsoft is the world's largest software maker by revenue" - is a useful insight about the company, but our system would not be able to correctly modify that to fit any other company.

In contrast, by selecting sentences from the web about the given company, the DD approach ensures that the resulting description will be both grammatical and relevant. It also results in a wider variety of expressions and a greater number of sentences. However, it can include nonessential facts that appear in a variety of different web venues. It is not surprising, therefore, that the hybrid approach performs better than either by itself.

While in this paper we focus on company descriptions, the system can be adapted to generate descriptions for other entities (e.g. Persons, Products) by updating the seed datasets for both approaches (to reflect the important facts for the desired descriptions) and retuning for best accuracy.

Acknowledgment

This research was supported in part by an IBM Shared University Research grant provided to the Computer Science Department of Columbia University.

References

Eugene Agichtein and Luis Gravano. 2000. Snowball: Extracting relations from large plain-text collections. In *Proceedings of the fifth ACM conference on Digital libraries*, pages 85–94. ACM.

Gabor Angeli, Percy Liang, and Dan Klein. 2010. A simple domain-independent probabilistic approach to generation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 502–512, Stroudsburg, PA, USA. Association for Computational Linguistics.

Or Biran and Kathleen McKeown. 2015. Discourse planning with an n-gram model of relations. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1973–1977, Lisbon, Portugal, September. Association for Computational Linguistics.

Sasha Blair-Goldensohn, Kathleen R. McKeown, and And rew Hazen Schlaikjer. 2003. Defscriber: a hy-

brid system for definitional qa. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 462–462.

Nadjet Bouayad-Agha, Gerard Casamayor, and Leo Wanner. 2011. Content selection from an ontology-based knowledge base for the generation of football summaries. In *Proceedings of the 13th European Workshop on Natural Language Generation*, ENLG '11, pages 72–81, Stroudsburg, PA, USA. Association for Computational Linguistics.

John Conroy, Judith Schlesinger, and Dianne O'Leary. 2006. Topic-focused multi-document summarization using an approximate or acle score. In *Proceedings of ACL*.

Hal Daumé III and Daniel Marcu. 2005. Bayesian multi-document summarization at mse. In *Proceedings of the Workshop on Multilingual Summarization Eva luation (MSE)*, Ann Arbor, MI, June 29.

Noemie Elhadad and Kathleen R. Mckeown. 2001. Towards generating patient specific summaries of medical articles. In *In Proceedings of NAACL-2001 Workshop Automatic*.

Michael Elhadad. 1991. *FUF: The Universal Unifier User Manual ; Version 5.0*. Department of Computer Science, Columbia University.

Güneş Erkan and Dragomir R. Radev. 2004. Lexrank: Graph-based centrality as salience in text summarization. *Journal of Artificial Intelligence Research (JAIR)*.

Elena Filatova and John Prager. 2005. Tell me what you do and i'll tell you what you are: Learning occupation-related activities for biographies. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 113–120, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dimitra Gkatzia, Helen F. Hastie, and Oliver Lemon. 2014. Finding middle ground? multi-objective natural language generation from time-series data. In Gosse Bouma and Yannick Parmentier, editors, *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden*, pages 210–214. The Association for Computer Linguistics.

Ravi Kondadadi, Blake Howald, and Frank Schilder. 2013. A statistical nlg framework for aggregated planning and realization. In *ACL (1)*, pages 1406–1415. The Association for Computer Linguistics.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 228–231. Association for Computational Linguistics.

Christian M.I.M. Matthiessen and John A. Bateman. 1991. Text generation and systemic-functional linguistics: experiences from english and japanese.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into texts. In *Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain.

Jahna Otterbacher, Gunes Erkan, and Dragomir R. Radev. 2005. Using random walks for question-focused sentence retrieval. In *Proceedings of HLT-EMNLP*.

Ehud Reiter and Robert Dale. 1997. Building applied natural language generation systems. *Nat. Lang. Eng.*, 3(1):57–87, March.

Christina Sauper and Regina Barzilay. 2009. Automatically generating wikipedia articles: A structure-aware approach. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, pages 208–216, Stroudsburg, PA, USA. Association for Computational Linguistics.

B. Schiffman, Inderjeet. Mani, and K. Concepcion. 2001. Producing biographical summaries: Combining linguistic knowledge with corpus statistics. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL-EACL 2001)*, Toulouse, France, July.

William Yang Wang, Kapil Thadani, and Kathleen McKeown. 2011. Identifying event descriptions using co-training with online news su mmaries. In *Proceedings of IJNLP*, Chiang-Mai, Thailand, November.

Ralph M. Weischedel, Jinxi Xu, and Ana Licuanan. 2004. A hybrid approach to answering biographical questions. In Mark T. Maybury, editor, *New Directions in Question Answering*, pages 59–70. AAAI Press.

Michael White. 2014. Towards surface realization with ccgs induced from dependencies. In *Proceedings of the 8th International Natural Language Generation Conference (INLG)*, pages 147–151, Philadelphia, Pennsylvania, U.S.A., June. Association for Computational Linguistics.