Domain Adaptable Semantic Clustering in Statistical NLG

Blake Howald, Ravikumar Kondadadi and Frank Schilder
Thomson Reuters, Research & Development
610 Opperman Drive, Eagan, MN 55123
firstname.lastname@thomsonreuters.com

Abstract

We present a hybrid natural language generation system that utilizes Discourse Representation Structures (DRSs) for statistically learning syntactic templates from a given domain of discourse in sentence "micro" planning. In particular, given a training corpus of target texts, we extract semantic predicates and domain general tags from each sentence and then organize the sentences using supervised clustering to represent the "conceptual meaning" of the corpus. The sentences, additionally tagged with domain specific information (determined separately), are reduced to templates. We use a SVM ranking model trained on a subset of the corpus to determine the optimal template during generation. The combination of the conceptual unit, a set of ranked syntactic templates, and a given set of information, constrains output selection and yields acceptable texts. Our system is evaluated with automatic, non–expert crowdsourced and expert evaluation metrics and, for generated *weather, financial* and *biography* texts, falls within acceptable ranges. Consequently, we argue that our DRS driven statistical and template–based method is robust and domain adaptable as, while content will be dictated by a target domain of discourse, significant investments in sentence planning can be minimized without sacrificing performance.

1 Introduction

In this paper, we propose a sentence (or "micro") planning system that can quickly adapt to new domains provided a corpus of sentences from the target domain is supplied. First, all sentences from the corpus are parsed and a semantic representation is generated. We used predicate and domain general named entities from Discourse Representation Structures (DRSs) derived by *Boxer*, a robust analysis tool that creates DRSs from text (Bos (2008)). Second, the sentences are automatically clustered by their conceptual meaning with a *k*-means clustering algorithm and then manually reviewed for consistency and purity. Third, named entity and domain specific content tagging creates banks of templates (syntactic representations) associated with the respective cluster (a "conceptual unit"). Finally, a ranking algorithm is used to train a ranker that determines the optimal template at a given point in the generated discourse given various features based on the conceptual units and the text derived so far.

Our system generates sentences from templates given a semantic representation as part of a larger Natural Language Generation ("NLG") system for three domains: *financial, biography* and *weather* (from the SumTime-Meteo corpus (Reiter et al. (2005))). NLG is traditionally seen as a multistage process whereby decisions are made on the type of text to be generated (communicative goal); entities, events and relationships that express the content of that text; and forging grammatical constructions with the content into a "natural" sounding text. These stages are articulated in a variety of architectures - for example, Bateman and Zock summarize NLG as follows: (1) Macro Planning creating a document plan; (2) Micro Planning sentence planning; (3) Surface Realization concatenating the information from (1-2) into coherent and grammatical text; and (4) Physical Presentation document layout considerations (formatting, titles, etc.) (Bateman and Zock (2003)). Each one of these stages can have several subtasks and vary considerably in terms of complexity (*see generally,* McKeown (1985); Hovy (1993); Reiter and Dale (2000)). However, in general, some abstract representation is developed in (1-2) and (3-4) deal with translating the abstraction to natural language largely through either rule–based or statistical approaches.

Significant human investments often need to be made to create systems from scratch. But while these systems may perform very well for a specific domain, extending to alternative domains may require starting over. Statistical approaches can streamline some human investment, but domain adaptability remains a concern. Finding the appropriate balance between investing in input and achieving an appropriate level of evaluated acceptance of the output, let alone whether or not the approach is adaptable, can be problematic. More abstracted representations may require more rules to process and generate acceptable texts while less abstract representations may require less rules but more investment in human resources. When evaluated, we find that our system produces texts that fall within acceptable ranges for automatic metrics (BLEU and METEOR), non-expert crowdsourced evaluations via CrowdFlower and expert evaluations of the *biography* domain (based on similar evaluation comparisons for other NLG systems).

Basile and Bos suggest that DRSs provide an appropriate form of abstraction for NLG tasks (Basile and Bos (2011)). The reason being that DRSs provide deep semantic content in the form of named entities, relationships between entities, identity relations and logical implications (e.g. negation, scope) all of which have a straight forward mapping to syntactic parses (e.g., within Combinatorial Categorial Grammar) and, in sum, provide a useable architecture to perform a myriad of NLG tasks. We adopt Discourse Representation Theory (Kamp and Reyle (1993)) as a starting point for our experiments for domain adaptable NLG. And while we only use a few features of the DRS in the current work, we anticipate that the logical representations in DRT can be useful for future work, as in improving the clustering of conceptual units in the training corpus, for example.

The main contributions of this paper are:

- A hybrid approach to sentence planning that combines a statistical system with a template-based system where templates are generated semi-automatically with minimal human review
- Domain adaptability is shown in three different domains (*financial, biography* and *weather*).
- Non-expert human evaluation is carried out by means of crowdsourcing. The evaluation provides scores for overall fluency of the generated text as well as sentence-level preferences between generated and original texts. These evaluations are supplemented by expert evaluations for the *biography* domain.

This article is structured as follows. Section 2 describes existing rule-based and statistical NLG approaches and domain adaptability. Section 3 explains our methodology; including DRSs and their use in clustering the three corpora and how the generated clusters are ranked and deployed in the generation of texts. Section 4 presents sample generated texts and the results of automatic and crowdsourced evaluations. Section 5 concludes with limitations and avenues of future research.

2 NLG: Templates, Rules and Statistics

This section discusses current approaches to NLG. We argue that a combination of a statistical approach and templates has an advantage over purely rule-based or statistical NLG systems.

Overall, NLG systems tend to be rule–based where some type of text is sought to be generated and different stores of data are manipulated to generate texts. The rules exist at all levels of the system from selecting content, to choosing a grammatical output to post-processing constraints (e.g. sentence aggregation and pronoun generation). For example, the SUMTIME-METEO project (Reiter et al. (2005)) generates weather forecasts from numerical data. First, the numerical data is analyzed, then decisions are made about what content to convey based on the analysis and how to grammatically represent the content at the document and sentence level. These decisions are implemented by hand crafted rules with input from multiple experts. Hence, rule–based systems come with a potentially high development cost due to the necessity of domain experts and system developers creating the rules.[1]

[1] Anja Belz references a personal communication with Ehud Reiter and Somayajulu Sripada where 12 person months where spent on the SUMTIME-METEO microplanner and realizer alone (Belz (2007)).

Statistical NLG systems, on the other hand, look to bypass or minimize extensive construction of rules by using corpus data to "learn" rules for one or more components of an NLG system (Langkilde and Knight (1998)). Alternative generations are then created from the rules and a decision model governs which alternative to choose at a given point in a generated discourse. For example, the *p*CRU system, which also generates weather texts from numerical data, starts with a small number of relations that are trained on a corpus (Belz (2007)). Other statistical systems such as the SPaRKy (Stent et al. (2004)) for generating restaurant recommendations uses a ranking algorithm for training rules for sentence generation. Statistical systems have less of a reliance on human input, but they require robust training data and it is harder to control the output – often leading to texts that are shorter, less natural and possible ungrammatical (but *see e.g.*, van Deemter et al. (2005)).

Our system relys on both statistical and template–based approaches. We first statistically learn the semantic structure of a given domain of discourse which is then used to produce templates for our system (combining the Micro Planning and Surface Realization stages). Next, to pick the best template, we train a ranker which ranks the different sentence templates (the SPaRKy system that also employs a ranking algorithm, but it ranks different *rules* rather than the *sentences*). This combination avoids pitfalls stemming from a statistical model generating the input for a realizer (also avoiding the need for an extensive grammar) and, in contrast to some systems which rely only in part on statistical learning (e.g., for template selection but not for generating underlying semantic structures (Galley et al. (2001))), we find that our approach is not only efficient in terms of processing and generating data, but also highly adaptable to different domains with minimized human involvement.

3 Methodology

In order to generate the different templates, it is necessary to rely on some formalism to capture the semantics of a given training corpus. Reducing the training corpus to semantic expressions works to ensure that use of human experts would be minimized and flexibility in domain adaptability could be preserved while not compromising the quality of the generated texts. To this end, we utilized *Boxer* which relies on a combination of CCG parsing, part–of–speech tagging and a store of lexical semantic representations from the CCGbank (Hockenmaier and Steedman (2005)) to create the structures. Each DRS is a combination of domain general named entities (DATE, PERSON, etc.) and predicates (typically content words, but also shallow semantic categories such as AGENT and EVENT) which are related by different relational elements (typically function words) (*in, by*). For our system, we extract only those words and categories marked as predicates and the domain general entity tags. To illustrate, consider (1):

(1) a. The consensus recommendation for the financial services peergroup is a buy.
b. T. Rowe Price led the group last quarter with a 3.6% average per share price increase.
c. The increase is projected to level off during the current quarter.

The predicate and domain general entity information created by *Boxer* for (1) is as follows:

(2) a. CONSENSUS | RECOMMENDATION | EVENT | SERVICE | PEERGROUP | BUY | ...
b. COMPANY | LEAD | DATE | SHARE | EVENT | AVERAGE | INCREASE | ...
c. INCREASE | EVENT | PROJECT | OFF | DATE | ...

The DRS-based predicates and domain general entities in (2) provide a lexical semantic representation of the sentence which captures the conceptual meaning of the sentence. Our assumption is that each grouping of DRS-based predicates represents the semantic "concept" of the sentence. The highly abstracted representation that does not utilize, for example, the relational information between the predicates, is a good starting point for grouping sentences and creating clusters (via *k*-means, discussed below in Section 3.1) by semantic concept. In viewing each sentence in a training corpus as such (indicated with an identifier ("*CuId*")), and a document as a sequence of "conceptual units" associated with templates and a store of predetermined information (domain specific tagging), we can categorize sentences by concept and create an organized bank of syntactic representations. For example, consider (3) (assuming, for the sake of presentation, that each utterance in (1) conveys a separate conceptual units):

(3) a. $\{CuId : 001\}$
Information: **industry**: financial services peergroup; **recommendation**: buy
b. $\{CuId : 002\}$
Information: **company**: T.Rowe Price; **time**: last quarter; **value**: 3.6%;
industry: the group; **financial**: average per share price; **movement**: increase
c. $\{CuId : 003\}$
Information: **movement**: increase, level off; **time**: the current quarter

The associated template representation (assigned to sentence in (1)) would be as follows:

(4) a. $\{CuId : 001\}$: The consensus recommendation for the **[industry]** is a **[recommendation]**.
b. $\{CuId : 002\}$: **[company]** led **[industry]** **[time]** with a **[value]** **[financial]** **[movement]**.
c. $\{CuId : 003\}$: The **[movement]** is projected to **[movement]** during **[time]**.

For domain adaptability in NLG, the key is to find a method that allows for the extraction of the appropriate level of semantics to be useable for generation across different corpora. The level of semantics can be relatively course or fine grained, weighed against a number of relevant factors (e.g., the communicative goal and the selection of content). The selection of content for our system is relatively fixed and is based on domain specific (not discussed here) and general tagging (e.g., COMPANY, DATE, PERSON from *Boxer* or other open source tools). Domain specific tags were not considered in the extraction of predicates from our training corpora. The following example from the *biography* domain illustrates the types of semantic content extracted for purposes of clustering the semantics of different training corpora.

(5) *Training Sentence*
a. Mr. Mitsutaka Kambe has been serving as Managing Director of the 77 Bank, Ltd.
since June 27, 2008.
b. Earlier in his career, he was Director of Market Sales, Director of Fund Securities and
Manager of Tokyo Branch in the Bank.
c. He holds a Bachelor's in finance from USC and a MBA from UCLA.
Conceptual Meaning
d. SERVING | MANAGING | DIRECTOR | PERSON | COMPANY | DATE | ...
e. EARLY | CAREER | DIRECTOR | MARKET | SALES | MANAGER | ...
f. HOLDS | BACHELOR | FINANCE | MBA | HOLD | EVENT | ...
Content Mapping
g. $\{CuId : 004\}$
Information: **person**: Mr. Mitsutaka Kambe; **title**: Managing Director;
company: 77 Bank, Ltd.; **date**: June 27, 2008
h. $\{CuId : 005\}$
Information: **person**: he; **title**: Director of Market Sales, Director of Fund Securities,
Manager; **organization**: Tokyo Branch; **company**: the Bank
i. $\{CuId : 006\}$
Information: **person**: he; **degree**: Bachelor's, MBA; **subject**: finance; **institution**: USC;
UCLA
Templates
j. $\{CuId : 004\}$: **[person]** has been serving as **[title]** of the **[company]** since **[date]**.
k. $\{CuId : 005\}$: Earlier in his career, **[person]** was **[title]**, **[title]** and **[title]** of **[organization]**
in **[company]**.
l. $\{CuId : 006\}$: **[person]** holds a **[degree]** in **[subject]** from **[institution]** and a **[degree]** from
[institution].

As shown in (4-5), predicate and domain general information from *Boxer* captures significant variability in the different domains of discourse which becomes less problematic with our approach than compared with, for example, rule–based sentence planning. This is with the proviso that a sufficiently sized and variable training corpus is available. Example generations for each domain are included in (6).

(6) *Financial*

a. First quarter profit per share for Brown-Forman Corporation expected to be $0.91 per share by analysts.

b. Brown-Forman Corporation July first quarter profits will be below that previously estimated by Wall Street with a range between $0.89 and $0.93 per share and a projected mean per share of $0.91 per share.

c. The consensus recommendation is Hold.

d. The recommendations made by ten analysts evaluating the company include one Strong Buy, one Buy, six Hold and two Underperform.

e. The average consensus recommendation for the Distillers peer group is a Hold.

Biography

f. Mr. Satomi Mitsuzaki has been serving as Managing Director of Mizuho Bank since June 27, 2008.

g. He was previously Director of Regional Compliance of Kyoto Branch.

h. He is a former Managing Executive Officer and Chief Executive Officer of new Industrial Finance Business Group in Mitsubishi Corporation.

Weather

i. Complex low from southern Norway will drift slowly nne to the Lofoten Islands by early tomorrow.

j. A ridge will persist to the west of British Isles for Saturday with a series of weak fronts moving east across the North Sea.

k. A front will move ene across the northern North Sea Saturday.

Because of the nature of our statistical plus template–based approach, it was not necessary to utilize all that *Boxer* has to offer. We only used predicates, which, for all intense and purposes, could be captured with content words, and domain general entity tagging. However, there are several additional aspects of *Boxer* which may prove useful such as exploiting the relation information, rhetorical relations and drawing further inferences based on the logical structure of the DRS are left to future work.

In sum, for our system, given some training sentence clustered on relatively simple semantics, coupled with domain specific tagging, templates can easily be generated and organized in a logical manner. With a large enough training corpus, there would be multiple templates (cf. Table 1) within each *CuId* and the one selected for generation would be statistically learned. The next section provides more detail about the data and clustering of semantic information in the creation of conceptual units and template banks from which the selection model generates text.

3.1 Data and Clustering

As indicated in Table 1, the *financial* domain includes 1067 machine generated texts from a commercially available NLG system covering mutual fund performance reports (n=162) and broker recommendations (n=905) from a commercially available NLG system, ranging from 1 to 21 segments (period ended sentences). The *biography* domain includes 1150 human generated texts focused on corporate office biographies, ranging from 3-17 segments. The *weather* domain includes 1045 human generated weather reports for offshore oil rigs from the SUMTIME-METEO corpus (Reiter et al. (2005)).

For each domain, the corpus was processed with *Boxer* and those items identified as predicates and named entity tags by the system were extracted. Each sentence then, represented as string of predicates and domain general tags, was clustered using *k*–means (in the WEKA toolkit (Witten and Frank (2005))) with *k* set to 50 for the *financial* domain and 100 for the *biography* and *weather* domains. The resulting clusters were manually checked to determine consistency - i.e., that all strings of predicates and

Table 1: Data and Semantic Cluster Distribution.

	Financial	*Biography*	*Weather*
Texts	1067	1150	1045
Conceptual Units	38	19	9
Templates	1379	2836	2749
Average Template/CU (Range)	36 (6–230)	236 (7–666)	305 (6–800)

tags assigned to a cluster conveyed the same or similar concept.[2] Clusters can be thought of as groups of most common words, for example the "recommend" cluster in the *financial* domain included RECOMMEND, CONSENSUS, COMPANY, the "current position" cluster in the *biography* domain included PERSON, POSITION, COMPANY, JOIN, DATE, and the "ridge" cluster in the *weather* domain included RIDGE, PRESSURE, DIRECTION.

The *biography* and *weather* domains, despite being human generated, are semantically less interesting (19 and 9 conceptual units respectively) but exhibit significantly more variability – 236 and 305 average number of templates per conceptual unit as compared to 36 for the *financial* domain (which is machine generated). The end result of the semantic preprocessing (along with domain specific entity tagging) is a training corpus reduced to templates (cf. 4,5j-l) organized by semantic concept. We use a ranking model to select a template corresponding to a semantic concept.

3.2 Ranking Model

For each conceptual unit, we rank all the matching templates and select the best ranked template. In order to train a ranking model, we do a 70/30 split of the data for training and testing. We represent each training document as a series of conceptual units along with the input information. For each conceptual unit, we first filter out all the non-matching templates by entity type and number - selecting only those templates that match the type of domain specific tagging present in the data and also have the same number of entities for each entity type. We rank the remaining templates based on the Levenshtein (Levenshtein (1966)) edit distance from the gold template (Template extracted from the original sentence in the training document). Additionally, several features are extracted for the top 20 ranked templates (to ease processing time) and are used in building the model: (1) N-grams: Word n-grams extracted from the template. We used 1-3 grams; and (2) Length: Normalized length of the input template. We used a ranking support vector machine (Joachims (2002)) with a linear kernel to train a model and each feature in the model will have an associated weight.

During testing, the system is presented with a sequence of conceptual units and the input data associated with each conceptual unit. All the templates associated with the conceptual units are extracted from the template bank and are filtered according to the filtering criteria used in the training phase. For each of the remaining templates, the model weights are applied to compute a score and the highest scored template is selected for generation. This embodiment constitutes the *system* generations. For the purpose of evaluation, we compared the *system* generations against the *original* texts and texts created without the ranking model - where any template associated with a conceptual unit is selected at random (rather than based on score) after applying the filter (*random* generations). The next section discusses the generated texts and a series of automatic and human (non-expert crowdsourced and expert) evaluations of the texts.

[2] To this end, we initialized *k* to an arbitraily large value to facilitate collapsing of similar clusters during manual verification. We assume this to be an easier task than reassigning individual sentences from existing clusters. As indicated in Table 1, this proved useful as the most semantically varied domain turned out to be the *financial* domain with 38 clusters (each cluster corresponds to a different conceptual unit).

4 Experimental Results

Table 2 provides generation comparisons for the system (*_Sys*), random (*_Rand*) and the original (*_Orig*) text from each domain. The variability of the generated texts ranges from a close similarity to the original text to slightly shorter, which, as mentioned in Section 2, is not an uncommon (Belz and Reiter (2006)), but not necessarily detrimental, observation for NLG systems (van Deemter et al. (2005)). The generated sentences can be equally informative and semantically similar to the original texts (e.g., the *financial* sentences in Table 2). The generated sentences can also be less informative, but semantically similar to the original texts (e.g., leaving out "manager" in *Bio_Sys*). However, there can be a fair amount of gradient semantic variation (e.g., moving northeast *to* a location vs. moving northeast *across* a location in *Weather_Sys* and "Director of Sales Planning" vs. "Director of Sales" in *Bio_Rand*).

Table 2: Example Texts.

System	**Text**
Fin_Orig	Funds in Small-Cap Growth category increase for week.
Fin_Sys	Small-Cap Growth funds increase for week.
Fin_Rand	Small-Cap Growth category funds increase for week.
Weather_Orig	Another weak cold front will move ne to Cornwall by later Friday.
Weather_Sys	Another weak cold front will move ne to Cornwall during Friday.
Weather_Rand	Another weak cold front from ne through the Cornwall will remain slow moving.
Bio_Orig	He previously served as Director of Sales Planning and Manager of Loan Center.
Bio_Sys	He previously served as Director of Sales in Loan Center of the Company.
Bio_Rand	He previously served as Director of Sales of the Company.

Some semantic differences are introduced in our system despite generating grammatical sentences. For example, "remain slow moving" (*Weather_Rand*) is not indicated in the original text. These types of differences are more common for *random* rather than *system* generations. However, the ultimate impact of these and other changes is best understood through a comparative evaluation of the texts with automatic and human evaluations.

4.1 Evaluations and Discussion

We evaluate our NLG system with automatic and human metrics and the correlations between them. The human evaluations can (and, in some circumstances, must be) performed by both non-experts and experts. We provide non-expert crowdsourced evaluations to determine grammatical, informative and semantic appropriateness and the same evaluations by several experts in biography generation.

The automatic metrics used here are BLEU–4 (Papineni et al. (2002)) and METEOR (v.1.3) (Denkowski and Lavie (2011)) and originate from machine translation research. BLEU–4 measures the degree of 4-gram overlap between documents. METEOR uses a unigram weighted *f*–score less a penalty based on chunking dissimilarity. We also calculated an error rate as an exact match between strings of a document. Table 3 provides the automatic evaluations of *financial, biography* and *weather* domains for both *random* and *system* for all of the testing documents in each domain (*financial* (367); *weather* (209); *biography* (350)).[3]

For each domain, the general trend is that *random* exhibits a higher error rate and lower BLEU–4 and METEOR scores as compared to *system*. This suggests that the *system* is more informative than the *random* text. However, scores for the *financial* domain exhibit a smaller difference compared to *weather* and *biography*. Further, the BLEU–4 and METEOR scores are very similar. This is arguably related to the fact that the average number of templates is significantly lower for the *financial* disourses than the *weather* and *biography* domains. That is to say, there is a greater chance of the *random* system selecting

[3] If comparing originals, the Error Rate would equal 0 and BLEU–4 and METEOR would equal 1.

Table 3: Automatic Metric Evaluations of *Biography*, *Financial* and *Weather Domains*.

Metric	*Bio_Rand*	*Bio_Sys*	*Fin_Rand*	*Fin_Sys*	*Weather_Rand*	*Weather_Sys*
Error Rate	0.815	0.350	0.571	0.477	0.996	0.698
BLEU–4	0.174	0.750	0.524	0.577	0.057	0.469
METEOR	0.198	0.520	0.409	0.386	0.256	0.436

the same template as *system*. So, from an automatic metric standpoint, applying model weights increases "performance" of the generation (based on coarse content overlap). However, human evaluations of the texts are necessary to confirm and augment what the automatic metrics indicate.

Two sets of crowdsourced human evaluation tasks (run on CrowdFlower) were constructed to compare against automatic metrics: (1) an understandability evaluation of the entire text on a three-point scale: **Fluent** = no grammatical or informative barriers; **Understandable** = some grammatical or informative barriers; **Disfluent** = significant grammatical or informative barriers; and (2) a sentence–level preference between sentence pairs (e.g., "Do you prefer Sentence A (from *original*) or the corresponding Sentence B (from *random/system*)"). 100 different texts and sentence pairs for *system*, *random* and the *original* texts from each domain were selected at random. Figure 1 presents the text understanding task and Figure 2 presents the sentence preference task (The aggregate percentage agreement for the text–understandability is .682 and .841 for the sentence–preference tasks based on four judgments per text and sentence pair).[4]

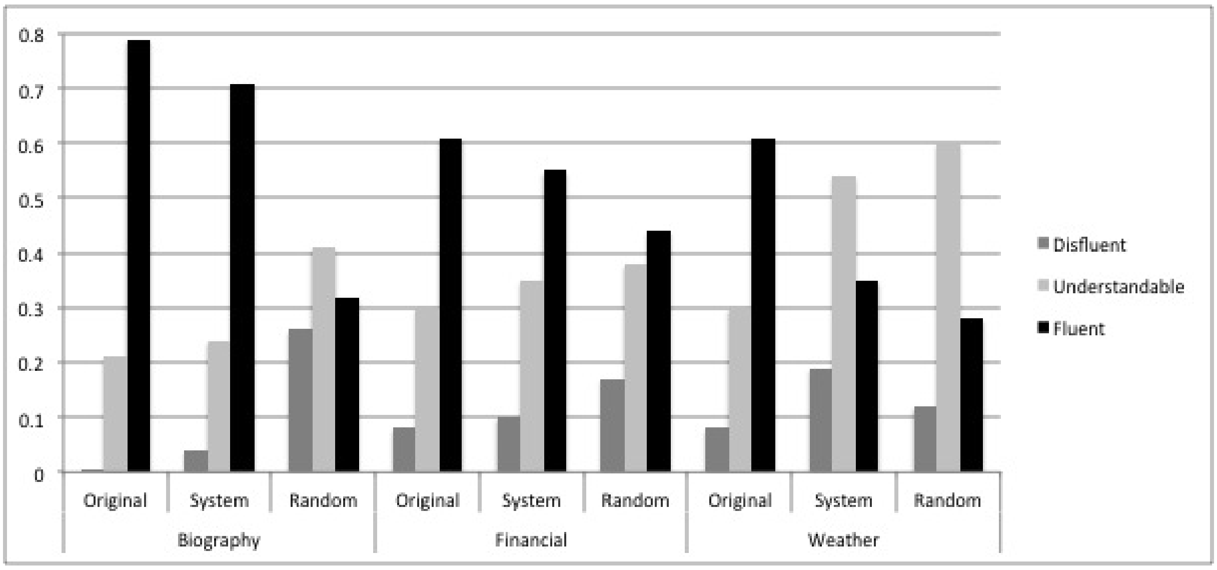

Figure 1: Human Text–Understandability Evaluations.

In all cases, the *original* texts in each domain demonstrate the highest comparative **fluency** and the lowest comparative **disfluency**. Further, the *system* texts demonstrate the highest **fluency** and the lowest **disfluency** compared to the *random* texts. However, the difference between the *system* and *random* for the *financial* and *weather* domains are fairly close whereas the differences for the *biography* domain is much greater. This makes sense as the *biography* domain is human generated and exhibits a high amount of variability. Given that the *weather* domain is also human generated and exhibits more variability compared to the *financial* domain, but they read more like the *financial* domain because of their narrow geographic and subject matter vernacular.

[4] Over 100 native English speakers contributed, each one restricted to providing no more than 50 responses and only after they sucessfully answered 4 "gold data" questions correctly. We also omitted those evaluators with a disproportionately high response rate. No other data was collected on the contributors (although geographic data (country, region, city) and ip addresses were available). Radio buttons were separated from the text to prevent click bias.

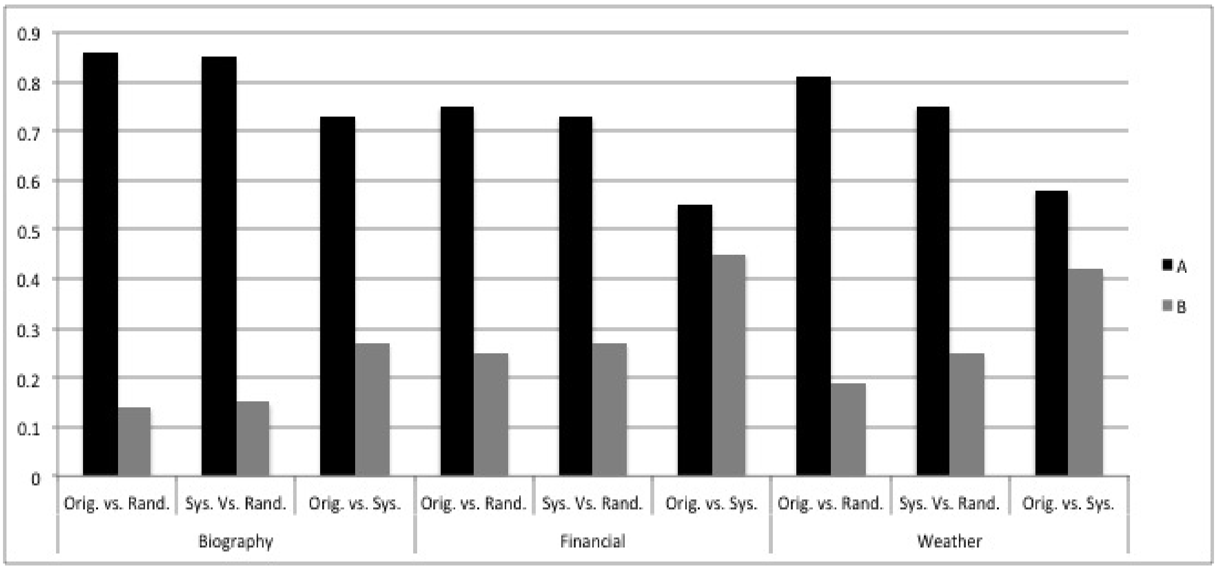

Figure 2: Human Sentence–Preference Evaluations.

Similar trends are demonstrated in the sentence preferences (Figure 2). In all cases, the *original* and *system* sentences are preferred to *random*. The *original* sentences are also preferred to *system* sentences, but the difference is very close for the *financial* and *weather* domains. This indicates that, at the sentence level, our *system* is performing similar to the *original* texts.

As indicated in Table 4, Pearson Correlation, based on 300 documents (100 from each domain), between the automatic metrics are high with the appropriate direction (e.g., error rate correlates negatively with BLEU–4 and METEOR scores, which correlate positively with each other). The human ratings - a consolidated score (**Fluent** = 1, **Understandable** = .66, **Disfluent** = .33) averaged over four raters per document - behave similar to the BLEU–4 and METEOR automatic metrics, but much less stong. There is more variability captured in the human judgments as compared to the automatic metrics which are both stricter and more consistent.

Table 4: Human–Automatic Pearson Correlation ($p \leq .0001$)).

	Error Rate	BLEU–4	METEOR	**Human**
Error Rate	1	-.719	-.715	-.406
BLEU–4		1	.827	.520
METEOR			1	.490
Human				1

Extreme cases aside, there is no exact formula for translating automatic and human evaluations to a true estimation for how the generated texts are performing. It is a relative determination at best and, in all actuality, deference is paid to the human evaluations. Human understandability of the texts is key.

We were able to perform expert evaluation of the *biography* domain. Three experts journalists, who write short biographies for news archives, performed the same two non–expert crowdsourced tasks. For the text evaluation, the experts rated both the *original* and *system* texts to be 100% **Fluent** (with the *random* texts following a similar distribution of non-expert ratings). For the sentence evaluations, the experts still preferred the *original* to the *system* sentences, but with an increase in preference for the *system* as compared to the non-experts - 27% preference by non-experts versus a 35% preference by experts. This trend is a reverse of what is reported for weather texts. For example, Belz and Reiter report a reduction in acceptability with experts as compared to non-experts (Belz and Reiter (2006)). This makes sense as the expert should be more discriminant based on experience. For the present texts, it could be the case that our system is capturing nuances of biography writing that experts are sensitive to. However, more critical expert feedback is required before saying more.

The performances that we present here are comparable to other rule–based and statistical systems. However, comparing systems can be problematic given the different goals and architectures. Nonetheless, the evaluations and generated texts indicate that we have been able to appropriately capture interesting and varied semantic structures.

5 Conclusions and Limitations

We have presented a hybrid statistical and template–based NLG system that generates acceptable texts for a number of different domains. Our experiments with both experts and non–experts indicate that the generated text is as good as the original text. From a resource standpoint, it is an attractive proposition to have a method to create NLG texts for a number of different subject matters with a minimal amount of development. The initial generation of the conceptual units and templates for the *financial* domain took two person weeks. This was reduced to two days for the *weather* and *biography* domains. Most of the development time was spent on domain specific tagging and model creation.

As compared to other NLG systems, there are several limitations to what we have presented here. First of all, our system assumes the document plan is given as an input; but this is not always necessarily true. In addition to the document plan, we also use domain specific tags from the original text. For example, we use phrases like *last quarter* as our input whereas a typical NLG system receives pure data like an exact date indicating the end of the quarter. It is the NLG system's responsibility to generate the corresponding referring expression appropriate for the current context. We are currently working on an extension of our framework that includes document planning and referring expression generation. This will also enable us to compare our system with existing state-of-the-art statistical NLG systems such as *p*CRU. We have not done expert evaluation for the *financial* and *weather* domains. While non-experts can provide useable judgments on the well–formedness of generated texts, evaluating the finer grained semantics of the text falls with the expert and will be included in future development. Finally, our system will only work with domains that have significant historical data. If only limited data is available, our system potentially cannot capture the variety of linguistic expressions used to express a semantic concept and will thus fail to avoid redundancy across texts.

Future work will focus on additional domains, and the integration of more discourse–level features into the model. Also, as we have only focused on a small part of what DRSs contain, deepening the semantics with the inclusion of relational elements may improve generation as well. We are in particular interested in utilizing the semantic representation for an improved clustering of conceptual units. As indicated in this article, attention to semantic structures is central to NLG and captures a large portion of the theoretical construction of such systems.

Acknowledgments

This research is made possible by Thomson Reuters Global Resources with particular thanks to Peter Pircher, David Rosenblatt and Jaclyn Sprtel for significant support. Thank you also to Khalid Al-Kofahi for encouragment and to Ben Hachey and three anonymous reviewers for critical feedback.

References

Basile, V. and J. Bos (2011). Towards generating text from discourse representation structures. In *Proceedings of the 13th European Workshop on Natural Language Generation (ENLG)*, pp. 145–150.

Bateman, J. and M. Zock (2003). Natural language generation. In R. Mitkov (Ed.), *Oxford Handbook of Computational Linguistics*, Research in Computational Semantics, pp. 284–304. Oxford University Press, Oxford.

Belz, A. (2007). Probabilistic generation of weather forecast texts. In *Proceedings of Human Language Technologies 2007: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT'07)*, pp. 164–171.

Belz, A. and E. Reiter (2006). Comparing automatic and human evaluation of NLG systems. In *Proceedings of the European Association for Computational Linguistics (EACL'06)*, pp. 313–320.

Bos, J. (2008). Wide-coverage semantic analysis with *Boxer*. In J. Bos and R. Delmonte (Eds.), *Semantics in Text Processing. STEP 2008 Conference Proceedings*, Volume 1 of *Research in Computational Semantics*, pp. 277–286. College Publications.

Denkowski, M. and A. Lavie (2011). Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation*, pp. 85–91.

Galley, M., E. Fosler-Lussier, and A. Potamianos (2001). Hybrid natural language generation for spoken dialogue systems. In *Proceedings of the 7th European Conference on Speech Communication and Technology*, pp. 1735–1738.

Hockenmaier, J. and M. Steedman (2005). CCGBANK: Users' manual. In *Department of Computer and Information Science Technical Report MS-CIS-05-09*. University of Pennsylvania, Philadelphia, PA.

Hovy, E. H. (1993). Automated discourse generation using discourse structure relations. *Artificial Intelligence 63*, 341–385.

Joachims, T. (2002). *Learning to Classify Text Using Support Vector Machines*. Kluwer.

Kamp, H. and U. Reyle (1993). *From Discourse to Logic; An Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and DRT.* Kluwer, Dordrecht.

Langkilde, I. and K. Knight (1998). Generation that exploits corpus-based statistical knowledge. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (ACL'98)*, pp. 704–710.

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady 10*, 707–710.

McKeown, K. R. (1985). *Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text*. Cambridge University Press.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, pp. 311–318.

Reiter, E. and R. Dale (2000). *Building Natural Language Generation Systems*. Cambridge University Press.

Reiter, E., S. Sripada, J. Hunter, and J. Yu (2005). Choosing words in computer-generated weather forecasts. *Artificial Intelligence 167*, 137–169.

Stent, A., R. Prasad, and M. Walker (2004). Trainable sentence planning from complex information presentations in spoken dialog systems. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL'04')*, pp. 79.

van Deemter, K., M. Theune, and E. Krahmer (2005). Real *vs.* template-based natural language generation: a false opposition? *Computational Linguistics 31*(1), 15–24.

Witten, I. and E. Frank (2005). *Data Mining: Practical Machine Learning Techniques with Java Implementation (2nd Ed.).* Morgan Kaufmann, San Francisco, CA.