Comprehension Driven Document Planning in Natural Language Generation Systems

Craig Thomson, Ehud Reiter, and Somayajulu Sripada

Department of Computing Science, University of Aberdeen:
{c.thomson, e.reiter, yaji.sripada}@abdn.ac.uk

Abstract

This paper proposes an approach to NLG system design which focuses on generating output text which can be more easily processed by the reader. Ways in which cognitive theory might be combined with existing NLG techniques are discussed and two simple experiments in content ordering are presented.

1 Introduction

Document Planning is a difficult task for which there has been little research compared with other aspects of NLG. This is evident in the recent survey of NLG (Gatt and Krahmer, 2017) which devotes relatively little space to the problem. Existing approaches focus on human authored corpora as a gold standard, extracting relations from them in order to structure messages.

This paper proposes NLG system design with the guiding principle of producing output text which is optimal input for the human comprehension process. Comprehension in this context is the readers ability to process text. Construction-Integration (CI) theory (Kintsch, 1998) states that the cognitive process of reading is iterative. A highly interconnected knowledge graph based on both the text and the readers prior knowledge forms the readers mental model as text is processed.

A basic system is presented which orders messages using distributional semantics. The system uses a simple graph database to generate texts describing a products suitability for a task. An example generated text can be seen in Figure 1 with message ID's shown in parentheses.

> *The Nepal Extreme consists of an outsole, a rand, an upper and a lining (M0). It has a crampon rating which is required for mountaineering (M1). The outsole is stiff which is good (M3). The outsole and the rand consist of rubber which is durable, this suits mountaineering (M2,M4). The upper consists of synthetic leather and synthetic fabric which are both durable and water-resistant, this is good (M10,M8,M9,M11). The lining has insulation which suits mountaineering (M5). It also consists of Gore-Tex which is waterproof, this is required for mountaineering (M7). Gore-Tex is breathable as well which is good (M6).*

Figure 1: Generated Product Description (message ID's in parenthesis)

2 NLG Document Planning

Document Planning within NLG is most often defined as the tasks of Content Determination and Document Structuring (Reiter and Dale, 1999). The system must decide what to say, how to say it and in what order. Further, if there is an optimal way to convey the input data such that the reader better comprehends it or is moved to action by it, then text should be generated in this way.

Approaches to Document Structuring generally focus on the relationships between messages. Schema based approaches (McKeown, 1985) and Rhetorical Structure Theory (RST) (Mann and Thompson, 1988), (Hovy, 1993) offer methods for generating text driven by the relations between messages or groups of messages. Whilst they have some limited success it is difficult to generalize

Proceedings of The 11th International Natural Language Generation Conference, pages 371–380,
Tilburg, The Netherlands, November 5-8, 2018.

them across domains. They also require much manual work and each relation needs to be defined by annotators who often do not agree even within a domain.

Machine Learning has been investigated as a method for both Content Determination and Document Structuring (Lapata, 2006), (Barzilay and Lee, 2004), (Liang et al., 2009). Such models rely on existing corpora within the domain and often paired data-corpus. If the domain is changed, or if there is disagreement as to what the correct corpus should be within a domain then such approaches experience difficulty.

3 The Construction-Integration Model

RST and Schemata focus on the relations between messages. Focus mechanisms can be used to check if different messages contain identical subjects or objects, utilizing this information when ordering messages. Focus (Sidner, 1979), Centering Theory (Grosz et al., 1995), (Poesio et al., 2004) and Scripts (Schank and Abelson, 1975) all offer appealing models for how people read and process text. It has been difficult however to implement any of them in an NLG context, especially in a general fashion. Perhaps this is because structuring a narrative is a complex task with many variables, of which individual approaches might only address a subset.

With the CI model (Kintsch, 1998) argues that a text is not in itself sufficient to account for the meaning acquired when the reader processes it. The reader's prior knowledge and experience add to the mental model which is established and iteratively modified as the text is processed. Whilst relations play a role, it is primarily the argument concepts of a proposition which activate and filter relevant concepts in the reader's mental model. This has some experimental backing, such as the work of (Schwanenflugel and White, 1991) which found that word priming from previous discourse altered the processing of words in future paragraphs.

Long-term memory (LTM) is the complete set of a readers knowledge. We do not have clearly indexed and direct access to this knowledge, even if it is relevant in the current context. Short-term memory (STM) contains our immediate thoughts although it is limited to a small number of concepts. The capacity of STM has a long history of study (Murdock Jr, 1960) and whilst estimates vary, they are often in the range of 10-15 terms. This is orders of magnitude lower than the number of terms and inferred concepts present in even short narratives, making STM an unsuitable mechanism to explain our ability to comprehend text (Ericsson and Kintsch, 1995). Kintsch describes human comprehension as loosely analogous to a computer system. Data is stored in both STM (registers / cache) as well as LTM (a large but slower access storage device). Working memory (WM) is the processor in this analogy. When a reader has expert knowledge of a domain they are able to use STM as an index to LTM, allowing for increased cognitive ability using WM. Expert readers can then create a rich set of inferences in their mental model which can then be used to better comprehend the text.

Kintsch suggests propositions as a suitable first class concept for modeling meaning in language. A proposition in the context of CI theory is a predicate-argument schema. This is a simplified view of propositions as conceived in formal logics. These simple propositions form complex ones, which in turn can be used to generate the text. CI theory adds additional nodes to this network which are not present in the text. These can be thought of as inferences based on the context of the propositions. Kintsch calls these 'knowledge elaborations' and relations between them and the proposition nodes from the text are added to form a complex highly interconnected network in the readers mental model.

4 Operationalization of the CI Model

Whilst a graph representation of data contains all the input information required to generate a given text, CI theory suggests this only forms a subset of the complete model of comprehension held by the reader. A system based on CI would require some method of simulating knowledge elaborations. This additional information would allow for the simulation of inferences, with the complete mental model being a combination of the propositional representation of the text and these inferences.

Distributional Semantic methods such as word embeddings created with word2vec (Mikolov et al., 2013) can provide indication of some kinds of relatedness between terms. Large data sets, such as Wikipedia, could be used as general knowledge. Domain specific corpora could also

be used if available, either in place of or in some combination with general knowledge. By combining a graph representation of propositions with distributional semantics we would appear to have something which at a very basic level fits the CI model proposed by Kintsch. The angle between vector representations of arguments can be used to weight the system knowledge graph.

Knowledge elaborations are not added to the system knowledge base, this is not possible as there is no direct access to the users mental model. If distributional semantics can provide weights for the edges in our system graph, these can be considered when planning content. An assumption is made that when a path has short inter-message distances (lower angles between vectors), readers will be able to construct a richer interconnected network of propositions and knowledge elaborations than they would when inter-message distance is high.

Content would be selected from the system graph based on queries which return subsets of the graph. For example all paths between a start node and an end node. With the optional aid of content structuring rules, these subgraphs can be clustered and ordered based upon the edge weights as well as the connectivity of propositions, the latter being similar to implementations such as the WISHFUL system (Zukerman and McConachy, 1998). An optimization function would need to be implemented which finds the most appropriate representation of text given the graph data, input queries, distributional semantics and any imposed structural rules.

The bottom-up approach suggested here could be used in combination with existing top-down methods such as RST. A domain expert could inform the writer of the most important factors for a specific NLG system, providing an outline for the system. The comprehension driven techniques would then provide a sensible default where the narrative structure has not been defined. A convention-over-configuration approach.

5 System

The simple system used for exploratory investigation of NLG motivated by CI theory is outlined here. This is an early version of the system and further work is required to properly assess its capabilities. System input is in the form of messages which are extracted as paths from a knowledge graph. These messages are then ordered to form a Document Plan, before simple Micro Planning techniques are applied and the text is realized.

5.1 Vector Space Model

The Vector Space Model (VSM) was created using the Python Gensim implementation of Word2Vec. The corpus was stripped of all characters which were not within the alphabet for the given language. The corpus was lemmatized (using spaCy). The VSM is trained on English Wikipedia using Word2Vec. The training settings were skip-gram with 600 dimensions, a window of 5, negative sampling of 5 and all words with a lower total frequency than 5 were discarded.

Whilst Word2Vec has been used as a starting point, it is possible that models generated using systems such as GloVe (Pennington et al., 2014) and ELMo (Peters et al., 2018) would improve an NLG system which relies upon distributional semantics. Vector Space Specialisation (Mrksic et al., 2017) may also be useful.

5.1.1 System Input

Figure 2 shows nodes and relations from the graph database (Neo4j) for a Product which consists of Components, with each Component being made of Materials. All Products, Components and Materials (collectively Items) may have Attributes which have suitabilities for different Tasks. Items may also have a direct suitability for a Task.

The system will describe the suitability of a product for a task. The input to the system is an unordered list of proposition chains, with each proposition chain itself being an ordered list of proposition triples. All possible paths from the product to the task are extracted from the graph shown in Figure 2. Directionality of the edges is ignored at this stage. To generate the text shown in Figure 1 each unique path from the product (Nepal Extreme) to the task (mountaineering) is found and combined to form the list of proposition chains shown in Figure 3. In the special case where a task requires an attribute which is not present on the product or any of its child items, a chain is created to represent it. These proposition chains are messages in the NLG system.

5.1.2 Ordering to form the Document Plan

A vector representation for each message is calculated by combining the vectors for each argu-

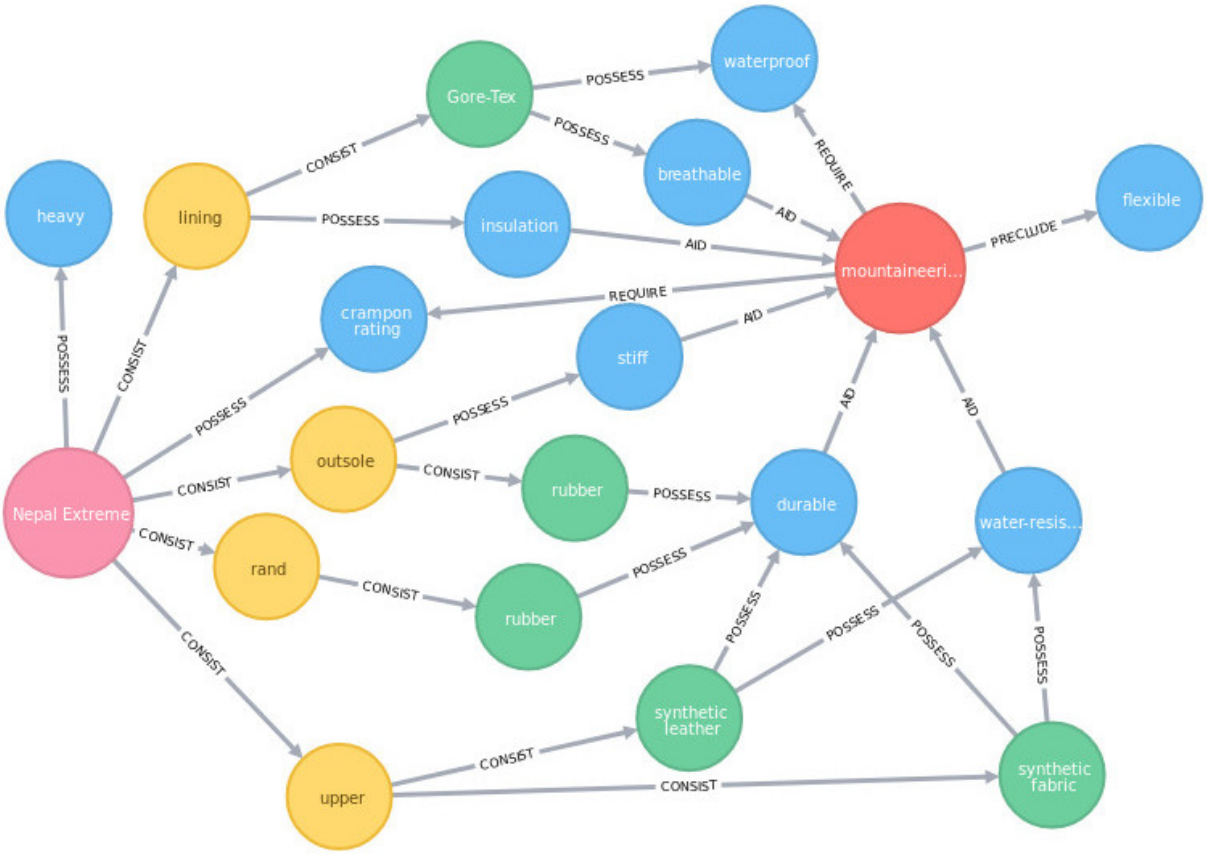

Figure 2: Product Composition, Attributes and Task.

ment. Vectors for individual arguments are taken from the VSM described in 5.1. The document plan is an ordered list of messages. To populate the document plan, messages are ordered using a greedy algorithm which minimizes inter-message distance. Inter-message distance is defined as the angle between the message vectors. The algorithm will stop once all input messages have been added to the document plan. The ordering of messages for the text example shown in Figure 1 can be seen, with inter-message distance in radians, in Figure 4.

5.2 Micro Planning

The focus of this paper is on content ordering at the Document Plan level. More advanced Micro Planning techniques could be used and would probably improve the quality of the text. The Aggregation and Referring Expression techniques (REG) used are far from state of the art and are meant only as a quick means to add variety and remove obvious repetition.

5.2.1 Aggregation

Functions were created to realize different patterns of proposition chains. The system iterates over the Document Plan, adding to a sublist of chains which are to be realized. If at any point the system determines it would be unable to realize this sublist, it reverts back to the last point at which it could and calls a suitable realizer function for the it. The system then starts from the chain which could not be realized, before continuing until all chains have been realized. It is always possible to realize a sublist of a single chain. Aggregation candidates can be seen on Figure 2 where paths diverge and re-converge, although the specific ordering means aggregation is not always possible.

Repeated propositions are removed at this stage. For example in Figure 3 the proposition triple at the beginning of both chains M3 and M4 describe how the ‘Nepal Extreme CONSIST outsole‘. This information was used for the purpose of ordering however it is only sent to be realized once, in the introductory sentence (M0) in Figure 1.

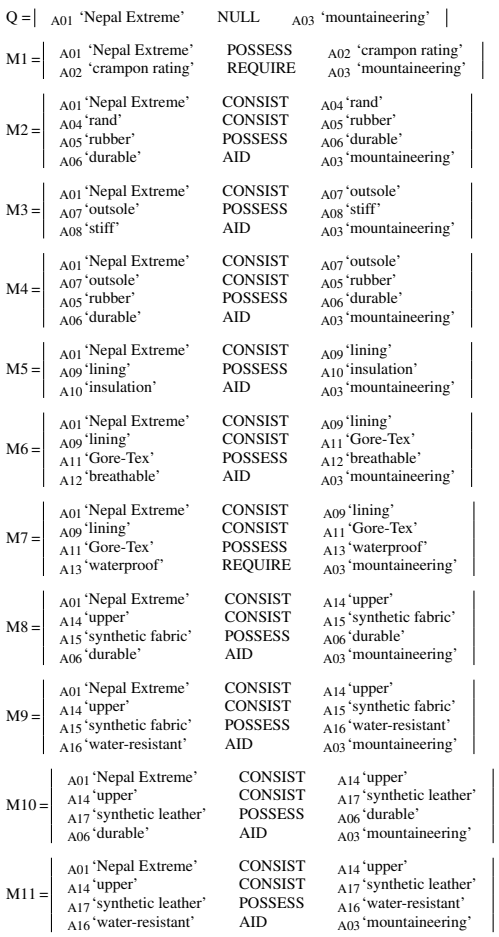

Figure 3: Input for text in Figure 1.

5.2.2 Referring Expression Generation

REG in the system is very simple. Pronouns are used only when the subject of the sentence is the same as that of the previous sentence. Whilst typically a Micro Planning task, this is done during realization, determined by the specific function which is called to realize the pattern of proposition chains.

5.3 Realization

Realization is performed using SimpleNLG (Gatt and Reiter, 2009). The realizer functions themselves use helper functions which construct commonly occurring patterns of text. An introductory sentence (labeled M0 in Figure 1) is included at the beginning of the output text detailing the product and its components. This is the only fixed ordering rule. The conjunction of components is realized in the order that the components would otherwise first be mentioned.

6 Experiments

6.1 Message Ordering

The first experiment evaluates the output of the simple product description system described in Section 5. The products within the system are all outdoor footwear. This domain was chosen because outdoor footwear can be broken down into a small number of components and attributes, then explained in broad terms. This would not hold true in a live system as there would be many ambiguity problems. It does however allow for a simple and contained preliminary test. The components of the product are parts of the boot/shoe such as upper, lining, rand and sole. Examples of materials are leather, suede and rubber. Attributes are most often adjectives such as durable or waterproof although they can also be concepts such as deep lugs. The tasks in this system are mountaineering, hiking and trail walking. With Item, Product, Component, Material, Attribute and Task being labels for nodes on the graph, the relations which are available are CONSIST, POSSESS, AID, HINDER, REQUIRE and PRECLUDE. These relationships can be seen on Figure 2. The experiment presented the below task descriptions to participants.

- Mountaineering - Walking and climbing in the mountains, often in the winter time when there is ice and snow.
- Hillwalking - Walking in the hills during every season except for winter. There may be some rough ground and it may be wet.
- Trail Walking - Walking on forest paths or other well kept trails. Usually in warmer weather although there may be some light rain.

In order to keep the system as simple as possible just four relations were used.

- AID - Attribute aids in the completion of the task, but is not essential. This is realized as 'is good for' or 'suits'.
- HINDER - Attribute hinders the completion of the task, but not to the point where it renders impossible. This is realized as 'is not good for' or 'does not suit'.

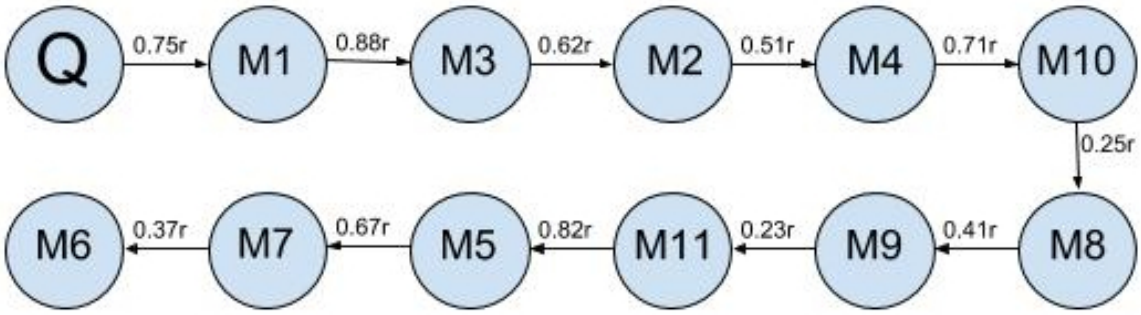

Figure 4: Ordering of Chains

- REQUIRE - Attribute is essential for the task. This is rendered as 'is essential for' or 'is required for'.

- PRECLUDE - Attribute precludes the item from the task. This was rendered as 'unsuitable for'. This was perhaps too close to 'does not suit'.

6.1.1 Experiment Setup

Participants were recruited for an online survey using email lists and social media. There were 37 respondents of which 28 stated English as their native language. Participants were shown 12 scenarios of which 6 were system generated narratives (like Figure 1) and 6 were hand crafted lists (like Figure 5 as might be seen on an online retail website. Each subset of 6 was equally divided such that participants received 3 product descriptions where the product was designed for the task (a match) and 3 descriptions where the product was not designed for the task (not a match). Participants were shown 4 statements and asked whether they agreed with each as it related to the current description using a five point Likert scale. The statements can be seen on Table 1

6.1.2 Hypothesis

Although this was exploratory research, the working hypothesis was that the ordering of the information would be evaluated as superior for the narrative descriptions, compared to list based descriptions. It was also suspected that narratives would outperform lists in the other categories although without a pilot experiment there was no real basis for this.

Attributes:

- Has a crampon rating.
- Upper - Made of synthetic leather and synthetic fabric. Is water-resistant and durable.
- Lining - Made of Gore-Tex. Is waterproof and breathable. Has insulation.
- Outsole - Made of durable rubber. Is stiff.
- Rand - Made of durable rubber.

Usefulness:

- Crampon rating and waterproof are essential for mountaineering.
- Water-resistant, durable, insulation, breathable and stiff are good for mountaineering.

Figure 5: List Based Description of Product Construction

6.1.3 Results and Evaluation

Table 1 shows the mean response for each question. Table 2 shows the results of the ANOVA test performed for each statement. The results show that several factors have an impact on the perceived quality of text. For all statements there was a statistically significant effect ($p < 0.005$) for whether the product being described was designed for the given task e.g. texts for mountaineering boots scored lower when being described for other tasks. There were also highly significant ($p < 3e\text{-}05$) effects when comparing narrative to list structure for all statements except S2 (The de-

scription explains the suitability of the product for the task). Lists were more highly rated overall although whether this would hold as the number of propositions increases is unclear and worthy of further examination.

There is a lot of ambiguity with the definitions used for the relations, Items, Attributes and Tasks. The line between Components and Attributes can be blurred, with 'deep lugs' being a good example. This could be a Component of the outsole or an Attribute of it. Ultimately the distinction has little impact on the system as it is the path from the Product to the Task which is important and this is unaffected by a change to the label of a node. Further, it is the name of the node which is used for word embeddings, not the label itself. The only impact for the final text is that the relation (and therefore verb) used will be CONSIST for a Component and POSSESS for an Attribute.

Where ambiguity is more of a problem is in the relations and the Tasks. There is a lot of overlap between mountaineering, hiking and trail walking. It is not always the case that these tasks are carried out in isolation. A mountaineer might walk over easier terrain using footwear which is suboptimal in order to approach a technical climb for which the footwear is essential.

6.2 Ordering Different Languages

A second experiment was conducted using the same system, the only difference being that argument concepts were translated to different languages for the purposes of ordering. Vector space models were trained for French and Spanish Wikipedia using almost identical settings as the English model (Section 5.1). The only change in the training pipeline was to allow additional characters such as accented vowels. Arguments were translated into French and Spanish by a single native speaker for each language. The domain was explained to each annotator and they were asked to translate arguments to the semantically closest word or phrase, with annotators having the ability to translate from a single term to a short phrase or vice versa.

6.2.1 Results and Evaluation

There were some difficulties in translation. Only one annotator was used per language. Any further work involving the definition of terms based on semantics, whether within English or to another language, should be done with multiple annotators and their agreement assessed.

Even with allowing annotators to translate into multiple terms, some did not directly translate. Hiking and Trail Walking were difficult to separate in French although ambiguity may exist in English as well. Trail Walking more often refers to walks at lower elevation on well maintained paths. Hiking includes walking off the path and on steeper, less stable terrain. A cursory Internet search for images based on these terms would appear to back this definition up, although exactly at which point Hiking becomes Trail Walking is ambiguous. The French annotator felt that 'randonnée' was the best term for both Hiking and Trail Walking although it was closer to Hiking. The phrase 'sentier de randonnée' is what Google translate returns for Trail Walking although this refers to the actual path which is walked upon, not the task. The concept could be expressed as a complex proposition although as this system only allows for simple lists of proposition triples, 'randonnée' was used for both tasks.

Table 3 shows the mean deviation of the ordering position of chains in Spanish and French when compared with the original English orderings. The overline indicates narratives where the stated product was not designed for the given task. It is difficult to evaluate the ordering based on translated proposition arguments as the English ordering makes for a poor gold standard. It is not clear if when using such a simple data source and such trivial propositions that there is a correct ordering.

Figure 6 shows the English realized text based on ordering using Spanish translations of graph nodes. The French example for the Nepal Extreme boot, used as an example throughout, was almost identical to the original English text in 1. Therefore, it has been omitted due to space restrictions.

7 Conclusion and Future Work

This paper describes a new approach to Document Planning based on the psychological model offered by Construction-Integration (CI) theory. It is interesting that CI suggests graph structure as a representation for human comprehension. Even if we cannot directly implement CI, the idea of manipulating graph data on the machine (speaker) end such that it might influence the 'graph data' on the human (hearer) end is worth pursuing. Investigation into this new approach is still in the early

	Statement	N	$\overline{N}$	L	$\overline{L}$
S1	The description is easy to read and understand	3.61	3.35	4.09	3.94
S2	The description explains the suitability of the product for the task	3.89	3.46	4.05	3.71
S3	The description is presented in a sensible order	3.63	3.29	4.05	3.75
S4	Overall, this is a good description	3.50	3.18	4.01	3.67

N : Narrative (matching Product)

L : List (matching Product)

$\overline{N}$: Narrative (non-matching Product)

$\overline{L}$: List (non-matching Product)

Table 1: Mean Response

	S1		S2		S3		S4	
Variable	F	p	F	p	F	p	F	p
Type	18.237	<0.001	2.135	0.145	18.43	<0.001	20.955	<0.001
Match	8.585	<0.01	22.685	<0.001	13.319	<0.001	15.003	<0.001
Participant	0.009	0.93	0.508	0.477	3.139	0.077	0.176	0.675
Product	1.001	0.44	1.503	0.129	1.376	0.183	1.615	0.093

Table 2: ANOVA Results

Task	French	Spanish	$\overline{French}$	$\overline{Spanish}$
Mountaineering	1.58	2.47	2.03	2.37
Hiking	2.60	2.27	2.96	1.67
Trail Walking	2.36	1.93	3.27	2.76

Table 3: Mean Order Variance per Language

stages and much remains to be done.

To fully test Comprehension Driven NLG, richer data sets and more comprehensive generation models will be required. Identifying and evaluating these are key prerequisites of future work. The qualitative evaluation of the first experiment presented in this paper only investigates the preferences of participants. *Evaluation of recall and deep understanding will also be required.* It is unclear as to whether the list based summary in its current form is a suitable gold standard to compare system generated narrative. Suitable methods of evaluating the Document Plan independently of the downstream system components will also be needed.

Existing approaches to Document Planning look at human authored corpora and attempt to construct narratives based upon patterns identified within them. This is either with hand crafted systems, ML/AI or a combination of the two. Whilst this corpus analysis is useful, it is a limitation of such methods that the text structure is insufficient to explain the comprehension process of reading

The Nepal Extreme consists of an upper, an outsole, a rand and a lining (M0). It has a crampon rating which is required for mountaineering (M1). The upper consists of synthetic fabric and synthetic leather which are both water-resistant and durable, this is good (M9,M11,M10,M8). The outsole and the rand consist of rubber which is durable, this suits mountaineering (M4,M2). It is also stiff which is good (M3). The lining consists of Gore-Tex which is waterproof (M7), this is required for mountaineering. Gore-Tex is breathable as well which suits mountaineering (M6). It has insulation which is good (M5).

Figure 6: Example Text (Spanish Order).

it.

Attempting to combine an NLG system's knowledge base with the mental knowledge base of the reader may appear highly impractical. Both however are processing systems, with the output from the former being the input to the latter. It therefore makes sense to optimize the writers output such that it can be more easily processed by the reader.

Future work will focus on identifying NLG techniques which generate output with this as a

primary consideration. It is the use of the human comprehension process itself, almost as a specialist node in a heterogeneous system, which will frame the research.

The most important task is designing an experiment which can demonstrate that either recall or deep understanding has been improved using an NLG system designed following comprehension principles. Work is in progress towards a system which will attempt to select optimal paths through a distributional semantical weighted proposition graph to explain a concept. *The system will look for paths, with smaller individual edge weights, rather than shorter paths which may be available but have greater inter-message distance.* It is hoped this will increase the chance that connected inferences are generated in the reader's mental model. Continued work on sentence ordering as discussed in this paper will be used to order the content.

As with many NLG systems, what has been discussed so far only operates on a small number of messages which at most would constitute a single paragraph or short communication format. *Document Planning is not however restricted to short texts and more research in the planning of long form documents is required.* Whilst this is a very complex task, paragraphs could be identified by clustering based on distributional semantics content. The most prominent propositions within paragraphs could be identified and used to generate top and tail statements, bridging paragraphs and even chapters. All of these techniques would first and foremost be comprehension driven.

Acknowledgments

This work is funded by the Engineering and Physical Sciences Research Council (EPSRC), which funds Craig Thomson under a National Productivity Investment Fund Doctoral Studentship (EP/R512412/1).

The authors would also like to thank Théo Morel and Alejandro Ramos Soto for translating the Argument Concepts to French and Spanish respectively.

References

Regina Barzilay and Lillian Lee. 2004. Catching the drift: Probabilistic content models, with applications to generation and summarization. *CoRR*, cs.CL/0405039.

K Anders Ericsson and Walter Kintsch. 1995. Long-term working memory. *Psychological review*, 102(2):211.

Albert Gatt and Emiel Krahmer. 2017. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *CoRR*, abs/1703.09902.

Albert Gatt and Ehud Reiter. 2009. Simplenlg: A realisation engine for practical applications. pages 90–93.

Barbara J. Grosz, Scott Weinstein, and Aravind K. Joshi. 1995. Centering: A framework for modeling the local coherence of discourse. *Comput. Linguist.*, 21(2):203–225.

Eduard H. Hovy. 1993. Automated discourse generation using discourse structure relations. *Artificial Intelligence*, 63(1):341 – 385.

Walter Kintsch. 1998. *Comprehension : a paradigm for cognition*, 1st edition. Cambridge University Press.

Mirella Lapata. 2006. Automatic evaluation of information ordering: Kendall's tau. *Comput. Linguist.*, 32(4):471–484.

Percy Liang, Michael I. Jordan, and Dan Klein. 2009. Learning semantic correspondences with less supervision. 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 91–99, Stroudsburg, PA, USA. Association for Computational Linguistics.

W.C. Mann and S.A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281. Cited By 820.

K.R. McKeown. 1985. Discourse strategies for generating natural-language text. *Artificial Intelligence*, 27(1):1–41. Cited By 112.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Nikola Mrksic, Ivan Vulic, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gasic, Anna Korhonen, and Steve J. Young. 2017. Semantic specialisation of distributional word vector spaces using monolingual and cross-lingual constraints. *CoRR*, abs/1706.00374.

Bennet B Murdock Jr. 1960. The immediate retention of unrelated words. *Journal of Experimental Psychology*, 60(4):222.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *CoRR*, abs/1802.05365.

Massimo Poesio, Rosemary Stevenson, Barbara Di Eugenio, and Janet Hitzeman. 2004. Centering: A parametric theory and its instantiations. *Comput. Linguist.*, 30(3):309–363.

Ehud Reiter and Robert Dale. 1999. *Building natural language generation systems*. Studies in natural language processing. Cambridge University Press, New York.

Roger C. Schank and Robert P. Abelson. 1975. Scripts, plans, and knowledge. In *Proceedings of the 4th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'75, pages 151–157, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Paula J. Schwanenflugel and Calvin R. White. 1991. The influence of paragraph information on the processing of upcoming words. *Reading Research Quarterly*, 26(2):160–177.

Candace L Sidner. 1979. Towards a computational theory of definite anaphora comprehension in english discourse. Technical report, Cambridge, MA, USA.

Ingrid Zukerman and Richard McConachy. 1998. An optimizing method for structuring inferentially linked discourse.