TOWARDS AUTOMATIC ROUTE DESCRIPTION UNIFICATION IN SPOKEN DIALOG SYSTEMS

Yulan Feng[1]*, Alan W Black*[1]*, Maxine Eskenazi*[1]

[1]Language Technologies Institute, Carnegie Mellon University, USA

2021 IEEE Spoken Language Technology Workshop (SLT) | 978-1-7281-7066-4/20/$31.00 ©2021 IEEE | DOI: 10.1109/SLT48900.2021.9383465

ABSTRACT

In telephone-based dialog navigation systems, scheduling and direction information are typically collected from routing APIs in text, and then delivered to users via speech. These systematic directions may be augmented with human descriptions to provide more accurate and personalized routes and cover broader user needs. However, manually collecting, transcribing, correcting, and rewriting human descriptions is time-consuming. Also its inconsistency with systematic directions can be confusing to users when delivered orally. This paper describes the construction of a pipeline to automate the route description unification process which also renders the resulting direction delivery more concise and consistent.

Index Terms— dialog systems, natural language generation, AI for good, assistive technologies

1. INTRODUCTION

Describing routes is essentially the simple task of providing directions from one place to another. Yet it can be a complicated research problem when it involves human language production and perception. As people have varying spatial reasoning abilities, when they cannot share a physical map, effective communication is a difficult problem. Psychologists have been studying the concept of cognitive maps in human navigation for decades [1, 2], and cognitive linguists have worked on spatial proposition classifications by analyzing collections of verbal routes [3]. Previous work on computer-generated route descriptions often relies on knowledge-based systems, and requires graphic user interfaces [4]. The route description generation and perception problem becomes even more difficult (and has rarely been studied) in the case of man-machine communication in spoken dialog systems. Route descriptions generated by systems like Google Maps and Bing Maps are concise and consistent but too general, many times missing details with reference to landmarks that are essential, for example, for disambiguation. Also, since these systems aim to provide global coverage, they can only provide directions from/to the main entrances of buildings, and the routes they provide in smaller local neighborhoods are often sub-optimal. On the other hand, experts such as local residents or information desk receptionists can provide more detailed and shorter paths in areas they are familiar with. However, as humans provide directions based on memory or a so-called "mental map", their route descriptions vary from time to time and person to person, and often contain colloquial and redundant information. This paper presents a pipeline to deliver concise and accurate route information to users in a spoken dialog system by combining human and system descriptions into a single navigation output. A new route direction dataset consisting of both synthesized audio and text forms of human and system descriptions is also presented.

2. BACKGROUND

2.1. Current System

GetGoing [5] is a spoken dialog system that provides trip planning information to users over a local phone line. The information is delivered to users through text-to-speech (TTS) after post-processing in the natural language generation (NLG) module. The system uses the Google Maps API as its backend for database lookup. It is able to correct potential automatic speech recognition (ASR) errors and resolve ambiguous locations. GetGoing also widely covers driving and public transit routes for Southwestern Pennsylvania. Since the system has been tailored to the needs of seniors who have limited access to smartphones or the internet, providing accurate and detailed directions to places such as hospitals is a top concern.

2.2. Application Scenario

Multiple major Pittsburgh hospitals are co-located on the UPMC Presbyterian (Presby) campus. The area is steeply sloped, giving the campus a complicated layout. Most buildings have multiple entrances on different levels, some being linked either by pedestrian bridges crossing over two streets or enclosed walkways connecting different floors of two buildings. Though Google Maps claims to have 99% coverage of the world, in our case, it fails to provide accurate walking directions for the Presby campus. Figure 1 shows an example of a bad route suggested by Google Maps. If a visitor wants to go from the bus stop (Fifth Ave at Halket St.) to the clinic (Benedum Geriatric Center), the route suggested

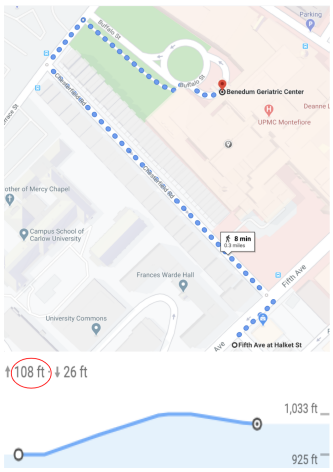

Fig. 1. An example bad route suggested by Google Maps

by Google Maps would involve 108 vertical feet of climbing, which is unnecessarily tiring for seniors with mobility difficulties or individuals using wheelchairs. If, instead of using Google Maps, visitors call the reception desk, they are told to access the clinic by entering the Montefiore building on Fifth Avenue, which is right across the street from the bus stop, and taking the elevator to the fourth floor. Incorporating human expert route descriptions in GetGoing is an essential complement to Google directions. GetGoing provides automated instructions to users who can ask for individual steps in the instructions to be repeated or for time to be given so that they can write something down. This makes it less intimidating to seniors who are anxious about talking to a real person over the phone [6]. It is also available 24/7.

3. RELATED WORK

3.1. Natural Language Generation

The goal of generating a unified route description can be considered to be a natural language generation (NLG) task. Recent advances in neural network methods have resulted in state-of-the-art results in multiple NLG tasks such as paraphrasing, summarization, and style transfer. However, these models generally require either parallel corpora [7] or massive amounts of data [8] for training. To the best of our knowledge, there is currently no large scale publicly available route description dataset. We therefore collected human route description data. With few sources of the human descriptions available, system-generated descriptions were used in training and in validating the sequence-to-sequence model (Transformer). Human descriptions were reserved for inference.

3.2. Route Description

Previous cognitive studies on urban route description [9, 3] show that people tend to describe routes using two major components: (1) landmarks that are reference points in the environment and (2) actions that are the instructions pedestrians are to follow. When both components are included, a recent spatial discourse study conducted on university campus settings [10] found that people tend to favor route descriptions that are short and more concise, and descriptions that are generated in groups. These features are especially preferred in communication situations where people need navigation assistance in an unfamiliar environment.

4. DATASET CONSTRUCTION

4.1. Systematic Descriptions

The systematic route descriptions were synthetically constructed using the Google Directions API. A list of street intersections in Pittsburgh was used to palliate ASR errors. To construct the system route description dataset, departure and arrival points were randomly chosen from this list. These points are then used to query the Google Directions API to obtain directions between the two locations. For example, for the route from Centre Ave and Penn Ave to N Euclid Ave and Penn Ave, the original output is "[Head northweston Penn Avetoward N Sheridan Ave,Slight rightto stay on Penn Ave,Take the crosswalk]". GetGoing breaks this down into sequences of steps, and post-processes each individual step into natural language, template, and abstract meaning representations. The resulting dataset has 85514 instances of steps. Some example instances can be found in Table 1.

Natural Language Representation Since the Google API directions are received in html format, the natural language representation of each step is formatted by removing all html elements, and resolving abbreviations such as Pl and Ct into complete English words (place and court respectively).

Template Extraction The steps returned by the Google Directions API mainly have 4 slots: (1) actions (e.g. turn, head, continue) with directions (*dir*), including both cardinal (north, south, east, and west) and relative (left, right, up, and down) directions, (2) road names and landmarks representing both the current place (*plc*) and the (3) destination (*dst*) to end the action, and (4) special instructions (*notes*) (e.g. "signs for S Side", "Take the stairs"). By replacing corresponding values in these slots with special *_unk* tokens, 40 route description templates were extracted.

Abstract Meaning Representation We then processed each step into an abstract meaning representation based on its natural language representation and the corresponding template. These representations were used as target sequences in the Transformer models proposed here.

4.2. Human Descriptions

Using a list of 12 common routes from/to bus stations, garages or parking lots, and main buildings around Presby, we interviewed and recorded 3 receptionists in different hospitals on the Presby campus with their consent. These recordings were automatically transcribed using Google's text to speech API followed by manual corrections and classification.

4.3. Data Augmentation

Since Google Maps does not cover indoor walking directions, 5 more templates were manually created to handle routes involving (1) vertical transportation such as elevators, escalators and stairs, (2) exit information including doors, buildings, and elevators and (3) indoor hallways. The data is thus augmented by filling in the templates with random place names from the same list that was used in dataset construction (4.1).

5. DESCRIPTION UNIFICATION

For a single trip, directly combining walking directions provided by human experts to proceeding driving or public transportation returned by routing systems can be inconsistent and thus confusing to users. An example can be found in Table 2.

5.1. Slot Recognition and Unification

To extract and unify key information from human descriptions, we implemented a slot recognizer[1] with a hybrid method combining rule-based and neural methods. The original parsing neural model was pretrained on OntoNotes 5.0 [12] with an imitation learning objective. This general-purpose model used trigram convolution layers [13] with residual connections as the encoder, followed by multi-layer perception to perform prediction. With a combination of bloom embedding strategy [14], layer normalization and maxout non-linearity, it achieved up to 85.43 Named Entity Recognition (NER) F score on the original evaluation data. We fine-tuned the model to predict the additional set of slots as defined in subsection 4.1, and further added rules to incorporate Part-of-speech (POS) tags and dependency relation information into the route slot recognition task. For example, for the description "take a second set of elevators that's in the eye and ear institute", the model first recognizes the following entity-text pairs: "cardinal:second, ele:elevators, organization:eye and ear institute", and this information is processed and combined to form the single place slot value "second set of elevators in eye and ear institute", where the redundant term "that's" is removed. This step helps to get rid of some superfluous information in human descriptions, and provides the flexibility of transferring the descriptions to other styles by either template filling or other NLG methods.

During the interview, all three experts were asked the same set of route questions to ensure the correctness and completeness of each individual route. Though their descriptions show a high rate of agreement in important information, there may occasionally be a detail missing or confused in one person's description in addition to different speaking styles. The routes were therefore automatically unified by combining information, and correcting slots after key information slots were extracted from each of their descriptions. For the example in Table 2, the key information "lower-level" in expert 3's description would thus be unified to "first floor" as mentioned by the other two experts based on majority vote. Further, expert 1's description ended with an ambiguous "go left or right", but the unified direction slot was filled with the solid value "right" since the other two experts both provided a concrete instruction "make a right".

This step makes future human description dataset construction easier and less time-consuming. Though it still requires collecting data from human experts, manual transcription, comparison and correction can be skipped. As we only need the key information from human descriptions, speech recognition errors made on the other parts of the recorded description will not affect the correctness of the final route description, and word error rates of the key information can be palliated with speech adaptation boosting.

5.2. Template Filling

For the template-filling approach, the templates extracted from systematic descriptions are categorized into different groups based on the set of slots they contain, regardless of the order. For example, templates (1) "continue straight onto *[plc]*", (2) "continue straight to stay on *[plc]*" and (3) "continue onto *[plc]*" all contain a single *[plc]* slot and thus belong to the same group. Then for each human description, with the slots key value pairs extracted in the entity recognizer model, a template with the same set of slots is randomly chosen, and filled with the corresponding values. For the example route in Table 2, given sentence "It can be accessed by going to the first floor of the garage", and the component "first floor of the garage" being tagged with slot *[plc]* by the recognizer, it is randomly filled to the "continue straight onto [plc]" template and thus description "Continue straight onto first floor of the garage" is generated by the template filling method.

5.3. Transformer Model

To both encode and generate varying-length texts, we adapted the Transformer translation model [15] [2] for the NLG task. The input of the model encoder in training is the abstract meaning representations of routes as can be seen in Table 1. At inference, input is a sequence representation of slot key and value pairs following the slot recognizer model (e.g.

[1] based on Spacy's NER system [11]

[2] followed implementation from fairseq

Head northwest on Penn Ave toward N Sheridan Ave.	Natural Language Representation
Head [dir_unk] on [plc_unk] toward [dst_unk].	Extracted Template
__start_dir__ northwest __end_dir__ __start_plc__ Penn Ave __end_plc__ __start_dst__ N Sheridan Ave __end_dst__	Meaning Representation
Slight left to stay on Greenfield Ave.	Natural Language Representation
Slight [dir_unk] to stay on [plc_unk].	Extracted Template
__start_dir__ left __end_dir__ __start_plc__ Greenfield Ave __end_plc__	Meaning Representation
Take the elevator to third floor of Children's Hospital	Natural Language Representation
Take the [ele_unk] to [plc_unk]	Extracted Template
__start_ele__ elevator __end_ele__ __start_plc__ third floor of Children's Hospital __end_plc__	Meaning Representation

Table 1. An example of different representations of single steps, where the third step was created by data augmentation

Google Map	Head southeast on Darragh Street toward Victoria St. Turn left onto Victoria St. Destination will be on the left.
Expert 1	It can be accessed by going to the **first floor of the garage**. And on the **first floor** a person will take a **second set of elevators that's in the eye and ear institute**. It will go to the **third floor** of the **eye and ear institute** and they can go **left or right**.
Expert 2	When you get out of the **garage** you'll have to go down to the **first floor** and then get the **elevator** up to the **third floor** and get off the elevator and make a **right**.
Expert 3	There is a **lower-level** lot you can get to the **Presby Garage** from there by taking the **elevator** up to the **third floor** and then getting off on the third floor and you'll see a sign that says **eye and ear**, you would make a **right** into that.
Template-filled	Continue straight onto **first floor of the garage**. Take the **second set of elevators in the eye and ear institute** to **third floor**. Turn **right** after exit **elevator**.
Transformer-generated	On **first floor of garage** take the **second set of elevators in the eye and ear institute**. On **elevator** take the **elevator** to **third floor**. Turn **right** after exit **elevator**, and you will be on **third floor**.

Table 2. Google map's instruction and different experts' descriptions of the following step.

"__start_plc__ first floor of the garage __end_plc__" for human description "It can be accessed by going to the first floor of the garage."). As in translation tasks, the decoder predicts the next word conditioned on the encoder output and the previous tokens in the generated sequence. Training target data is the natural language representation of the corresponding route as in Table 1. The encoder and decoder are trained together to optimize the label smoothing cross entropy of the training data. Since the order of slot positions does not matter in the direction description, we trained the transformer model with position embeddings removed. To handle new location names in the unseen test set, we use a negative unknown token penalty, and copy over the unknown tokens in the source sequence to the target sequence. The Transformer model is assessed using two sets of metrics, (1) the standard BLEU4 [16] score as in machine translation tasks, and (2) slot matching after description translation. The model achieves a BLEU4 score of 89.84, and a slot match accuracy of 98.45% on the validation set. Due to the same resource of training data and limited templates used by Google Maps, the descriptions generated by the Transformer are essentially very similar to the ones generated by the template-filling method.

6. HUMAN EVALUATION

In order to deliver the resulting directions in GetGoing, human users' preferences and system usefulness are a main concern. The qualities of route descriptions were assessed by synthesizing each of them into a single speech audio file. For each route, there were three versions: human description, template generation, and Transformer generation. To eliminate the effect of an individual person's accent etc, human description recordings were also resynthesized using the transcribed text. All three types of utterances were synthesized using the same synthetic voice as the one in the current version of GetGoing, with a normal speaking rate and no additional prosody features (sample recordings can be found in supplemental materials accompanying this submission). Using this set of synthesized utterances, we carried out two studies on Amazon Mechanical Turk (AMT): a memory test

and a preference test. Each AMT worker had a HIT approval rate (the proportion of completed tasks that are approved by requesters) greater than 95% and supplied information about whether they were native English speakers, and their age group.

6.1. Memory Test

The memory test included 3 random audio files per HIT, and at most one version was included for each route. Workers could hear each file once, and they then responded to the open-ended question "What key information do you remember? For example, what turn should you take, and where should you make the turn? Which floor should you go to? Which elevator should you take?". To evaluate information retention, this question appeared after the worker had finished listening to the first audio file. There were 275 workers, and each worker was allowed to submit this HIT only once. Noisy answers like "yes", "nothing" were removed in later analysis.

6.2. Preference Test

The preference test had 5 pairs of different routes per HIT, each pair consisting of 2 versions of the same route randomly chosen out of 3 possibilities (Transformer, template, human). There were 100 workers who could submit multiple HITs. For each pair, workers were asked an explicit comparison: "Which description do you prefer?", with 3 options "A", "B", and "No Preference", followed by an open-ended question "Why?". They were instructed to put "NA" if they could not think of any particular reason, and their responses were used for qualitative analysis. We use Fleiss' kappa [17] to evaluate inter-rater reliability.

6.3. Quantitative Results

Memory test: We quantified workers' responses to the memory test by counting the slots they remembered compared to the manually-labeled slots of a single route. Incorrectly remembered slots (e.g. third floor remembered as first floor) were discarded, and compound location slots (third floor of UPMC Presby) were counted as two separate slots. Table 3 shows the overall slot match accuracy results. The template version outperformed the other two versions overall in this memory test. In particular, for the first question of a HIT, where workers did not know what the exact task was and thus might have payed less attention to memorization, workers who listened to the template version had the most stable memory performance.

Preference test: With the order of versions randomized within each pair, Table 4 shows the weighted preference of each choice after outlier workers (i.e. response always different from the majority) were removed. The N-1 Chi-Square test, as recommended by Campbell [18] for preference testing analysis, shows that significantly more people preferred the template version to both the human (p=0.0013) and the Transformer version (p=0.0454). While this preference test is highly subjective (thus all 100 raters do not achieve overall agreement (kappa < 0.2)), each set of raters has fair agreement (kappa=0.4) on each individual HIT.

6.4. Qualitative Analysis

After removing all "NA" entries, the 234 preference reasons given in the open-ended why question in preference test were manually analyzed. Route description preferences can be grouped in five classes: *naturalness*, *brevity*, *clarity*, *memorability* and *details*. For all reasons given, with stopwords removed and all words lemmatized, some of the most frequently mentioned key words are "clear","concise","easy","understand" and "natural". While some reasons (16) advocate more elaborate and descriptive directions: "B is more helpful because it is more descriptive. ", more preference reasons (56) explicitly favored the shorter template versions: "B uses fewer words to say the same thing".

7. DISCUSSION

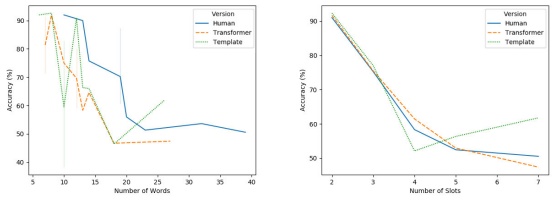

Fig. 2. Line plots of slot match accuracy against sentence length (left) and number of slots (right) in the memory task.

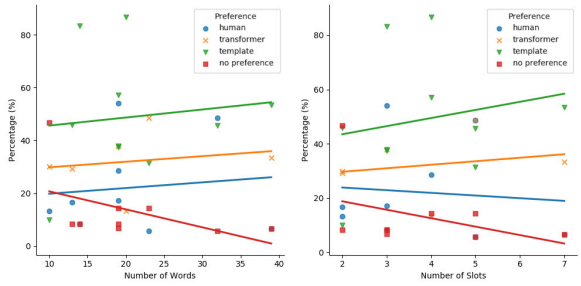

Fig. 3. Preference percentage against human description sentence length (left) and number of slots (right). Each point represents a version's preference per route.

.

Whether people's immediate recall ability is limited by the number of chunks [19, 20] to be recalled, or the length of the verbal list to be remembered [20] has long been debated

	Human		Template		Transformer		p-value		
	% (#samples)	std	% (#samples)	std	% (#samples)	std	h vs t	h vs s	t vs s
Overall	61.07(234)	0.31	65.96(294)	0.30	61.78(219)	0.31	0.1503	0.7391	0.2881
First only	57.49(86)	0.33	65.83(85)	0.31	59.75(83)	0.32	0.1157	0.4684	0.4082
Senior overall	66.67(47)	0.30	74.35(50)	0.27	75.64(44)	0.26	0.1299	0.1553	0.9412

Table 3. Slot match accuracy results for memory task. h,t,s denotes human, template, and Transformer respectively

Version	Human	Template	Trans.	Equal
Preference (%)	40.48	51.33	40.87	8.98

Table 4. Overall weighted preference results calculated by the number of times a version is chosen out of number of times that version is present in the pair. Equal denotes no preference, and Trans. denotes Transformer.

in cognitive science. In Figure 2, we show how workers' performance on the memory task is both affected by the number of words (length) and the number of information slots (chunks) in a single route description. The number of words affects memory performance less regularly on shorter sentences because, for example, the long location name "Eye and Ear Institute" was listed on the task page as an entity recognition hint. After close examination of individual routes, we see that workers listening to all three versions perform similarly (p=0.7) on directions with as few as two slots (first floor,garage). For example, human description "It can be accessed by going to the first floor of the garage", template description "continue straight onto first floor of the garage", and Transformer description "Continue onto first floor of the garage" had an average accuracy of 90%, 93%, and 91% respectively. However, the template version significantly (p=0.02) outperformed the human version on a more complicated route with 5 slots (lower level,elevator,third floor,Presby,Garage). The 22 workers who listened to the human description "There is a lower-level lot you can get to the Presby Garage from there by taking the elevator up to the third floor" on average remembered 51.30% of the key information, while the other 28 workers who listened to the template description correctly remembered 66.32%. The results thus show that delivering route descriptions using templates can help people continue to retain much information even as the number of words or chunks increases. It should be noted that while the Transformer version has lower results here, it is possible that other Transformer versions may perform better.

The scatter plots fitted with regression lines in Figure 3 suggest that while listeners may have no preference for a specific version when routes are short and simple, they tend to explicitly prefer more structured vocal deliveries of routes (template and Transformer versions) as the descriptions get longer, containing more information. The routes where raters' preferences are statistically significant ($p <$ 0.05) are the ones with more (4,5,7) slot information. This preference result on synthesized speech further confirms Denis' finding [10] on university campus directions that people tend to favor shorter route descriptions. The only exception is the human description "and then getting off on the third floor and walking halfway up the bridge and you'll see a sign that says eye and ear institute, you would make a right into that" compared to the template description "turn right after exiting on third floor, and you will be on bridge (signs for eye and ear institute)". While some raters still preferred the template version because "instruction is clear", the majority of workers (27/52) preferred the human version with reasons like "more detailed but not too complicated" and "it is intuitive". These reasons motivate us to work on balancing detail and brevity going forward.

Currently we can see the potential of Transformer model outperforming human description in the preference test. However, as a lot of new entity and indoor place names are not covered in training, they are unknown tokens to the model at inference and could lead to noisiness in the target sequence generation. This also shows the limit of automatic evaluation. Despite of the model's high performance on the dev set, a single mistake on the test set drastically affects human perception of the audio, especially for route memory purpose. So far we have tried with increasing unknown mapping threshold in training and tuning the unknown word penalty at inference time, and we will continue working on better copy mechanisms in the future. We will also work on incorporating template selection into the Transformer model for more controllable outputs.

8. CONCLUSIONS

This paper proposes to combine verbal human route descriptions and system descriptions through an automated pipeline. Human evaluations show that the combined directions delivered via audio are easier to remember and are preferred compared to original human descriptions.

9. ACKNOWLEDGEMENTS

This paper is supported by the Department of Transportation (DOT) under Mobility21 grant 69A3551747111. The views expressed in this paper do not necessarily reflect those of the DOT.

10. REFERENCES

[1] Sabine Timpf, GARY S. Volta, David W. Pollock, and Max J. Egenhofer, "A conceptual model of wayfinding using multiple levels of abstraction," in *Theories and Methods of Spatio-Temporal Reasoning in Geographic Space*, A. U. Frank, I. Campari, and U. Formentini, Eds., Berlin, Heidelberg, 1992, pp. 348–367, Springer Berlin Heidelberg.

[2] Barbara Tversky, "Cognitive maps, cognitive collages, and spatial mental models," in *Spatial Information Theory A Theoretical Basis for GIS*, Andrew U. Frank and Irene Campari, Eds., Berlin, Heidelberg, 1993, pp. 14–24, Springer Berlin Heidelberg.

[3] Michel Denis, "The description of routes: A cognitive approach to the production of spatial discourse," *Cahiers de Psychologie Cognitive*, vol. 16, pp. 409–458, 08 1997.

[4] Bernard Moulin, Driss Kettani, Benjamin Gauthier, and Walid Chaker, "Using object influence areas to quantitatively deal with neighborhood and perception in route descriptions," in *Advances in Artificial Intelligence*, Howard J. Hamilton, Ed., Berlin, Heidelberg, 2000, pp. 69–81, Springer Berlin Heidelberg.

[5] Shikib Mehri, Alan W Black, and Maxine Eskenazi, "Cmu getgoing: An understandable and memorable dialog system for seniors," *arXiv preprint arXiv:1909.01322*, 2019.

[6] Yusong Gao, Ang Li, Tingshao Zhu, Xiaoqian Liu, and Xingyun Liu, "How smartphone usage correlates with social anxiety and loneliness," *PeerJ*, vol. 4, pp. e2197, 07 2016.

[7] Regina Barzilay and Kathleen R. McKeown, "Extracting paraphrases from a parallel corpus," in *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, France, July 2001, pp. 50–57, Association for Computational Linguistics.

[8] Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola, "Style transfer from non-parallel text by cross-alignment," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., pp. 6830–6841. Curran Associates, Inc., 2017.

[9] Gérard Ligozat, "From language to motion, and back: Generating and using route descriptions," in *Natural Language Processing — NLP 2000*, Dimitris N. Christodoulakis, Ed., Berlin, Heidelberg, 2000, pp. 328–345, Springer Berlin Heidelberg.

[10] Daniel Marie-Paule and Michel Denis, "The production of route directions: Investigating conditions that favour conciseness in spatial discourse," *Applied Cognitive Psychology*, vol. 18, pp. 57 – 75, 01 2004.

[11] Matthew Honnibal and Mark Johnson, "An improved non-monotonic transition system for dependency parsing," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, September 2015, pp. 1373–1378, Association for Computational Linguistics.

[12] Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel, "Ontonotes: The 90in *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, USA, 2006, NAACL-Short '06, p. 57–60, Association for Computational Linguistics.

[13] Yann LeCun and Yoshua Bengio, *Convolutional Networks for Images, Speech, and Time Series*, p. 255–258, MIT Press, Cambridge, MA, USA, 1998.

[14] Burton H. Bloom, "Space/time trade-offs in hash coding with allowable errors," *Commun. ACM*, vol. 13, pp. 422–426, 1970.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., pp. 5998–6008. Curran Associates, Inc., 2017.

[16] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, USA, 2002, ACL '02, p. 311–318, Association for Computational Linguistics.

[17] JL Fleiss, "Measuring nominal scale agreement among many raters," *Psychological bulletin*, vol. 76, no. 5, pp. 378—382, November 1971.

[18] Ian Campbell, "Chi-squared and fisher–irwin tests of two-by-two tables with small sample recommendations," *Statistics in Medicine*, vol. 26, no. 19, pp. 3661–3675, 2007.

[19] Endel Tulving and Jeannette E Patkau, "Concurrent effects of contextual constraint and word frequency on immediate recall and learning of verbal material.," *Canadian Journal of Psychology/Revue canadienne de psychologie*, vol. 16, no. 2, pp. 83, 1962.

[20] Alan D Baddeley, Neil Thomson, and Mary Buchanan, "Word length and the structure of short-term memory," *Journal of verbal learning and verbal behavior*, vol. 14, no. 6, pp. 575–589, 1975.