Improved Text Generation using N-gram Statistics

Eder Miranda de Novais, Thiago Dias Tadeu, and Ivandré Paraboni

School of Arts, Sciences and Humanities, University of São Paulo (USP / EACH)
Av. Arlindo Bettio, 1000 - São Paulo, Brazil
{eder.novais,thiagoo,ivandre}@usp.br

Abstract. *In Natural Language Generation (NLG) systems, a general-purpose surface realisation module will usually require the underlying application to provide highly detailed input knowledge about the target sentence. As an attempt to reduce some of this complexity, in this paper we follow a traditional approach to NLG and present a number of experiments involving the use of n-gram language models as an aid to an otherwise rule-based text generation approach. By freeing the application from the burden of providing a linguistically- rich input specification, and also by taking some of the generation decisions away from the surface realisation module, we expect to make NLG techniques accessible to a wider range of potential applications.*

Key words: Text Generation, Surface Realisation, Language Modelling

1 Introduction

Natural Language Generation (NLG) systems enable non-linguistic data to be visualised as text reports. Systems of this kind often follow a pipelined architecture as proposed in [1], in which content messages are extracted from the domain data and are subsequently enriched with linguistic information, up to the point in which a full (and usually language-independent) abstract sentence specification is obtained. In the final stages of the NLG pipeline, sentence specifications are assigned content words and then linearised, in the task known as surface realisation [2, 3].

Surface realisation takes as an input a semantic representation of *what* to say, and determines *how* to say it, producing output surface strings in a given target language. A general-purpose surface realisation module will usually require a highly detailed, linguistically-motivated input specification to be provided by the underlying application. For instance, in order to generate a sentence as 'The foreign students will review this paper', a surface realisation system as [4] will normally expect as an input a feature structure as the following simplified example[1]:

[1] Some of the specified features in this example (e.g., action person) may be required in highly-inflected languages such as Portuguese, but this may not necessarily be the case in other languages, e.g., English.

```
sentence(
           voice=active,
           agent(   head=student, det=definite,
                    gender=masc, number=plural, premodifier=foreign),
           action(  head=review, tense=future,
                    mode=indicative, person=3p),
           patient( head=paper, det=demonstrative,
                    dettype=this, gender=masc, number=singular)
        )
```

In a wide-coverage surface realisation system, such rich representation is compulsory if we are to have full control over the output text. However, unless the underlying application is linguistically-oriented by design, it may be difficult to provide information in this level of detail for a number of reasons. First, this representation assumes that all content words are known in advance, that is, lexical choice has already been performed and the surface realisation module is set to use, e.g., the word 'student' (and not 'pupil' or 'learner' etc.) to represent the agent head concept.

Second, this representation also assumes that the agent property 'foreign' is to be realised in a pre-modifier position, which is not straightforward even when a single modifier is involved (compare to, e.g., 'the student (that was) mentioned', in which the participle modifier appears after the concept head.)

Finally, morphologically-rich languages require nouns, adjectives, participles and others to be gender- and number-inflected. Thus, unless the application is assumed to provide this kind of information as well, the surface realisation module must provide a full grammar for each target language under consideration.

These examples highlight the fact that either the application or the surface realisation module will have to implement - at a considerable cost - a number of language-dependent generation decisions. As an attempt to overcome some of these difficulties, in this paper we follow [5] and present a series of experiments involving the use of n-gram language models as an aid to an otherwise rule-based generation approach for the Brazilian Portuguese language. In doing so, we would like to free the application from the burden of providing a full input specification (hence making it easier to adapt), and also to take some of the generation decisions away from the surface realisation module (making it easier to deploy to a new target language.)

The reminder of this paper is organised as follows. Section 2 describes the surface realisation issues that are the main focus of our experiments and related work in the field. Section 3 presents each of the experiment settings and their results. Section 4 discusses our findings, and Section 5 presents our conclusions and future work.

2 Experiments overview

In what follows we use the surface realisation system described in [4] to investigate the role of n-gram language models as a means to simplify the system input

specification, and also as a possible substitute for some of the required grammar rules. To this end, three realisation-related problems will be taken as our working examples: lexical choice[2], ordering of nominal modifiers, and verb-complement agreement. The sentences to be generated are similar to those found in a small corpus of emails in an undergraduate project domain, conveying both questions made by students and replies sent by their project tutor.

In all experiments, our general strategy consists of leaving some of the input knowledge or rules unspecified, and using a trainable n-gram language model of Portuguese to guide the decision-making as proposed in [5]. The language models under consideration are standard 2-gram and 3-gram models[3] with back-off, built from a 40-million words corpus of Brazilian Portuguese newspapers articles. The NLG system itself, a standard template-based approach that takes feature structures as an input to produce word strings, will not be presently discussed - see [4] for details.

Our first research question concerns lexical choice, that is, the task of finding suitable words to represent each input concept given a potentially wide choice of synonyms [6]. The issue becomes particularly complex if, as pointed out in [7], we consider that sense-tagged corpora are rarely obtainable, and that the surface realisation of a given concept may have to be disambiguated not only among a potentially large number of synonyms, but among a large number of synsets in the first place.

In our experiments, two instances of lexical choice decision will be considered: those involving the realisation of the head constituents of noun phrases (NPs) and the head constituents of verb phrases (VPs.) Our goal in both cases is to investigate whether these choices can be left unspecified in the input data, leaving the decisions to be made with the aid of a n-gram language model. NP and VP head choices will be investigated separately in Experiments 1 and 2 in the next section, and also as a combined Experiment 3 which evaluates the limits of the n-gram approach to lexical choice in more complex decision-making.

Our second research question concerns the ordering of nominal modifiers, that is, the task of determining the linear order of realisation of adjectives, determiners, participles etc. attached to a noun head. The task is known to be considerably complex even for languages such as English, in which modifiers are mainly concentrated in the pre-nominal position (e.g., 'the small red book', but not '*the small book red'), and the correct orderings are not easily captured by grammar rules. Existing approaches tend to focus on the ordering of English pre-modifiers, which has been addressed as a classification problem [8, 9]. We notice however that modifiers in Romance languages may often appear in post-position as well, and that they may be easily combined with pre-modifiers in a single NP.

We will consider the ordering of noun modifiers in pre- and post-position, and also both simultaneously. Our goal in this case is to avoid the implemen-

[2] Lexical choice is often assumed to take place *before* surface realisation proper, as in [3]. However, our experiments are not committed to any particular NLG architecture.

[3] The choice for n-gram models of order 2 and 3 only was motivated by the amount of available training data.

tation of complex ordering constraints (or otherwise forcing the application to provide information in this level of detail) by using a n-gram model to select the appropriate configuration. This issue will be investigated in our Experiment 4.

Finally, the output sentences in a NLG system are of course expected to be grammatical, and a surface realisation module will normally implement a number of rules to enforce agreement and other linguistic constraints. Some of these constraints may however be difficult (or costly) to model unless a full grammar of each target language is considered. In [4], for instance, some long-distance dependencies are not fully supported by the surface realisation system, and the underlying application is expected to provide (or otherwise inherit from a pre-defined template) the correct gender and number attributes of certain sentence constituents.

We will address one particular kind of long-distance dependency that is not currently supported by the system described in [4], namely, the issue of verb and complement agreement. For instance, in 'The man who sold this house is very old', a morphologically-rich language such as Portuguese will require gender and number agreement between the subject (man) and the complement (old), which makes an interesting test case for the n-gram approach. We will consider those dependencies established in both active and passive voice, and our goal in both cases is once again to substitute language-dependent rules for a trainable statistical language model. These issues will be investigated in Experiment 5 and 6 described in the next section.

3 Evaluation

In this section we present the settings and results of Experiment 1-6 introduced in the previous section. Decisions made with the aid of language models of order 2 and 3 (called our Bigram and Trigram systems) will be compared to a number of baseline systems deemed relevant to each task. In all cases, the evaluation work will measure Edit-distance, NIST, BLEU and Accuracy scores obtained by each systems with respect to a manually-built reference set.

Briefly, Edit-distance is the traditional Levenshtein's distance (i.e., the number of insert, delete and substitute operations needed to make both strings identical.) Accuracy is taken to be the number of exact matches between the two strings, being assigned the value 1 if both strings are identical, or zero otherwise.[4] NIST [10] and BLEU [11] are widely-used evaluation metrics for Machine Translation systems based on n-gram counts[5].

3.1 Lexical Choice Preferences

The first experiment considers the lexical choice preferences of NP heads in subject position. To this end, 40 sentences were extracted from the corpus, making

[4] As an overly strict evaluation metric that expects exact string match, Accuracy scores are presented for illustration purposes only.

[5] BLEU scores range from 0 to 1, being 1 equal to 100% accuracy. NIST scores do not have an upper bound, but higher NIST scores stand for higher accuracy.

a Reference set. Next, multiple versions of each sentence were generated by making use of all possible surface forms for each NP head, taken from a thesaurus of Portuguese nouns. The thesaurus conveyed, on average, 3 alternatives for each NP found in the Reference data, and thus 120 alternative sentence realisations were produced out of the original 40-sentences set.

In order to select the most likely surface realisation form for each sentence, we consider the use of bigram and trigram language models compared against two baseline systems: a frequency-based approach that selects, for each concept, the most common word found in the corpus, and a strategy that simply chooses a random word out of each synset. The results are presented in Table 1. Recall that best results correspond to *lower* Edit-distance and *higher* Accuracy, NIST and BLEU scores.

Table 1. NP head lexical choice

System	Edit-distance	Accuracy	NIST	BLEU
Statistical 2-gram	2.88	0.60	9.048	0.955
Statistical 3-gram	1.73	0.75	9.151	0.969
Frequency-based	2.95	0.60	9.013	0.954
Random	4.70	0.40	8.842	0.928

Next, we consider the issue of lexical choice preferences as applied to VP head selection. The task in this case is considerably more complex than in the case of NP head selection given that (Portuguese) verbs tend to have a large number of synonyms: for instance, each verb synset in our data conveys an average of 10 synonyms.

Keeping all other sentence constituents unchanged, the 40 sentences in the Reference set used in Experiment 1 were re-generated, this time allowing only the main VP head constituents to vary according to the synonymous found in the thesaurus. As a result, 400 alternative sentences were produced. Following the same strategies from the previous experiment (Bigram, Trigram, Frequency-based and Random generation), the results were obtained as in Table 2.

Table 2. VP head lexical choice

System	Edit-distance	Accuracy	NIST	BLEU
Statistical 2-gram	4.98	0.13	8.731	0.863
Statistical 3-gram	3.88	0.30	8.877	0.893
Frequency-based	3.25	0.42	8.891	0.907
Random	5.38	0.13	8.659	0.852

Finally, in order to assess the limits of the n-gram approach to lexical choice, Experiment 1 and 2 were combined into a single task. In other words, the 40

Reference sentences were once again re-generated while attempting all possible NP and VP surface realisation forms available from the thesaurus, producing a set of 1164 sentences in total. The results of the Bigram, Trigram, Frequency-based and Random strategies are presented in Table 3.

Table 3. Combined NP-VP head lexical choice

System	Edit-distance	Accuracy	NIST	BLEU
Statistical 2-gram	8.25	0.08	8.375	0.829
Statistical 3-gram	6.05	0.25	8.607	0.868
Frequency-based	6.08	0.20	8.552	0.867
Random	7.88	0.05	8.338	0.822

3.2 Ordering Constraints

In a second class of experiments we investigated whether strict linguistic constraints (and not merely domain preferences) that are notoriously difficult to capture in the form of grammar rules could be enforced by the use of statistical language models. To this end, we take the issue of ordering of noun modifiers as our working example.

A Reference set of 40 sentences was manually built as follows. Each sentence conveyed a simple subject-verb-complement structure in the undergraduate project domain. NP heads in subject position conveyed one or two modifiers each, and their correct position was always fixed, that is, cases in which more than one reading were possible were avoided.

The ordering of modifiers was evenly distributed across the data: in cases of single modifier, half instances required a pre-modifier (e.g., 'the red book') and half required a post-modifier (e.g., 'the book mentioned (in section 8)'[6]); in the case of two simultaneous modifiers, their instances were evenly distributed in three categories: using two pre-modifiers (e.g., 'the small red book'), using two post-modifiers[7], and using one pre-modifier and one post-modifier.

In our corpus, it is often the case that a noun is modified by prepositional phrases (PPs), as in, e.g., 'the cover of the book', and since PPs widen the distance between sentence constituents, these expressions pose a considerable challenge to the use of n-gram models in surface realisation. Thus, regardless of the number or position of modifiers, in this experiment we consider two kinds of NPs: those conveying a single noun head, and those conveying a noun head attached to a prepositional phrase. The number of sentences using a single NP or an NP-PP attachment was also evenly distributed in the Reference set.

[6] Examples of post-modifier usage may seem more natural in our target language (Portuguese) as in, e.g., 'O livro citado'.

[7] Once again, these are commonly seen in Portuguese, as in 'o prazo máximo permitido' ('the latest deadline allowed'.)

Keeping all other constituents unchanged, the 40 sentences in the Reference set were generated considering all possible orderings of modifiers for the nouns in subject position. As a result, 132 alternative sentences were produced, with an average of 3.2 realisations for each Reference sentence.

Unlike previous Experiment 1-3, which focused on domain preferences that could in principle be gauged at by means of corpus analysis, the present experiment addresses strict linguistic constraints that the text generator has to adhere to, making a comparison against a frequency-based approach unsuitable. For this reason, in Experiment 4 we will limit our analysis to the results of the n-gram models as compared to a single baseline system that selects one of the possible modifier orderings at random. The results are shown in Table 4.

Table 4. Ordering of noun modifiers

System	Edit-distance	Accuracy	NIST	BLEU
Statistical 2-gram	5.15	0.65	7.244	0.814
Statistical 3-gram	5.35	0.63	7.201	0.795
Random	9.73	0.27	6.831	0.559

3.3 Agreement Constraints

Finally, we examined the role of n-gram language models in verb and complement agreement, a problem which becomes obviously more complex, from a n-gram perspective, as additional words are inserted within the agreeing constituents.

Our starting point was the Reference set built for the previous Experiment 4 (on modifier ordering constraints). In order to increase the scope of the present investigation, we considered not only the 40 original sentences, but also modified versions of them in which the distance between verb and complement was increased by inserting one or two adverbs. Thus, a Reference set of 120 sentences was built, divided in three groups conveying 0, 1 or 2 intermediate constituents.

The complement gender (male/female) and number (singular/plural) attributes were evenly distributed across the data, although we did not expect an effect on this. The following is an example of each group.

(a) The students are keen to learn.
(b) The students are not keen to learn.
(c) The students are not very keen to learn.

We used 40 * 3 = 120 sentences in a standard subject + verb + complement order in active voice, making the Reference set for Experiment 5. In addition to that, we designed a separate Experiment 6 using a Reference set in which the same 120 instances were re-written in passive voice. By breaking the most typical n-gram chains in this way, we expect to have made the agreement task

considerably more complex, allowing us to further assess the possible benefits and limitations of the n-gram-based approach.

In both Experiment 5 (agreement in active voice) and 6 (passive voice), we applied the same procedure adopted in the previous experiments. Keeping all other sentence constituents unchanged, our tests consisted once again of over-generating the Reference sentences, this time allowing only the gender and number of the complement term to vary. In doing so, four alternative surface realisations for each sentence were generated, producing 480 output sentences in Experiment 5 and another 480 output sentences in Experiment 6. The results produced according to the Bigram, Trigram and Random strategies are summarised in Table 5 and 6.

Table 5. Verb-complement agreement constraints (active voice)

System	Edit-distance	Accuracy	NIST	BLEU
Statistical 2-gram	0.08	0.93	7.557	0.980
Statistical 3-gram	0.08	0.93	7.557	0.980
Random	0.99	0.25	6.467	0.791

Table 6. Verb-complement agreement constraints (passive voice)

System	Edit-distance	Accuracy	NIST	BLEU
Statistical 2-gram	0.45	0.67	7.132	0.886
Statistical 3-gram	0.42	0.68	7.167	0.892
Random	0.95	0.30	6.473	0.754

4 Discussion

Experiment 1-3 on lexical choice preferences showed overall positive results for NPs, VPs and NP-VP combined. The task seems however more easily accomplished in the case of NP head lexical choice than in the case of VPs. This was to some extent to be expected as verbs in our data tend to have a much larger number of synonyms than nouns.

Regarding the NP head lexical choice task, the Trigram strategy (and, to a lesser extent, also Bigram) was superior to the others. By contrast, in the VP head lexical choice task, the frequency-based approach was best of all.

As for the combination of NP-VP lexical choice, all systems showed a relatively low performance, suggesting that n-gram models do not provide a suitable distinction between complex, structurally-dependent phenomena. Although in

this case the Trigram and Bigram models were still the best options, our results suggest that these issues should be addressed separately, or with the aid of a more powerful language model (e.g., taking syntactic information into account.)

Experiment 4 - on ordering constraints of noun modifiers - showed once again that the use of n-gram language models outperform the baseline system. In this case, the Bigram model was best of all, although the difference between Bigram and Trigram results was not statistically significant. Despite the presently higher Accuracy scores if compared to, e.g., those obtained in the previous experiments, we notice that establishing the ordering of NP modifiers was in some ways more difficult than performing lexical choice, as observed by the measured Edit-distance, NIST and BLEU scores.

Finally, regarding Experiment 5-6, we observe that the statistical models were both remarkably efficient in the verb-complement agreement task, with a small (but non-significant) advantage of Trigram over Bigram. More importantly, although not shown in the previous section, these results were found to be constant regardless of the distance between verb and complement[8], i.e., regardless of using zero, one or two intermediate words[9]. In addition to that, the experiments showed that the results for sentences in active voice were considerably superior to those obtained in passive voice, which may be explained by the fact that active voice sentences are much more frequent in the training data.

5 Conclusions

This paper presented a number of experiments to model typical surface realisation decisions with the aid of n-gram language models as a first step towards a simplified development and use of NLG resources. To this end, three tasks were considered: the lexical choice of NP and VP head constituents, the ordering of noun modifiers and verb-complement agreement.

Our preliminary results suggest that, as in [5], the use of n-gram statistics may indeed improve surface realisation of morphologically-rich languages such as Portuguese in a number of ways. In particular, the NP head lexical choice and verb-complement agreement tasks were successfully accomplished with the aid of language models, and that may indeed free the underlying application from the burden of providing this level of linguistic detail to the language generator.

By contrast, the VP head lexical choice was to some extent less successful. This may be partially explained both by the small size of the training data set and by the naivety of the language models under consideration.

As future work, we intend to make use of more sophisticated models that take into account not only n-gram statistics, but also morphological and structural properties of text such as those supported by factored language models [12].

[8] There was actually a small increase in NIST/BLEU scores as longer sentences were considered, which is to be expected given the nature of these n-gram-based metrics.

[9] Although we are presently unable to validate this claim, we assume that a sufficiently robust language model of higher order may in principle be able to cope with longer dependency relations in a similar fashion.

Acknowledgments. The authors acknowledge support by FAPESP and CNPq.

References

1. Reiter, E.: An Architecture for Data-to-Text Systems. In: European Natural Language Generation workshop (ENLG-2007), pp. 97-104 (2007)
2. McRoy, S., Channarukul, S., Ali, S.S.: An augmented template-based approach to text realization. Natural Language Engineering 9 (4) pp. 381-420. Cambridge University Press (2003)
3. Gatt, A., Reiter, E.: SimpleNLG: A realization engine for practical applications. In: European Natural Language Generation workshop (ENLG-2009) (2009)
4. Novais, E.M., Tadeu, T.D., Paraboni, I.: Text Generation for Brazilian Portuguese: the Surface Realization Task. NAACL-HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas, pp. 125-131 (2010)
5. Langkilde, I.: Forest-based statistical sentence generation. In: Proceedings of ANLP-NAACL'00, pp. 170-177 (2000)
6. Reiter, E., Sripada, S.: Human Variation and Lexical Choice. Computational Linguistics 28 (4) (2002)
7. Bangalore, S., Rambow, O.: Corpus-based lexical choice in natural language generation. In: 38th Meeting of the ACL, Hong Kong, pp. 464-471 (2000)
8. Malouf, R. : The order of prenominal adjectives in natural language generation. In: Proceedings of ACL-2000, Hong Kong (2000)
9. Mitchell, M.: Class-Based Ordering of Prenominal Modifiers. In: Proceedings of the 12th European Workshop on Natural Language Generation, Athens, pp. 50-57 (2009)
10. NIST: Automatic Evaluation of Machine Translation Quality using n-gram Co-occurrence Statistics. `www.nist.gov/speech/tests/mt/doc/ngram-study.pdf` (2002)
11. Papineni, S., Roukos, T., Ward, W., Zhu, W.: Bleu: a method for automatic evaluation of machine translation. In: ACL-2002, pp. 311-318 (2002)
12. Bilmes, J., Kirchhoff, K.: Factored Language Models and Generalized Parallel Backoff. In: Proceedings of HLT/NACCL, pp.4-6. (2003)